AUTOMATIC SPEECH RECOGNITION: EXPERIMENTS WITH A RECOGNISER USING LINGUISTIC STATISTICS

by

P. Denes

Technical Note No. 4
Contract No. AF 61(514)-1176 U.S. Air Force
September 1960

DEPARTMENT OF PHONETICS
University College London
England

AUTOMATIC SPEECH RECOGNITION: EXPERIMENTS WITH A RECOGNISER USING LINGUISTIC STATISTICS

P. DENES

Department of Phonetics
University College London
England

TECHNICAL NOTE NR. 4

Contract Nr. AF 61(514)-1176
September 1960

The research reported in this document has been sponsored in part by the Air Force Cambridge Research Center, Electronics Research Directorate, of the Air Research and Development Command, United States Air Force, through its European Office.

ABSTRACT

The problem of transmitting speech over communication channels with smaller information-carrying capacity than that of conventional telephone links is discussed.

Bandwidth compression systems uaing articulatory constraints (vocoders) are described and this is followed by a description of devices that analyse the speech sound wave in terms of linguistic units - machines performing this task are called automatic speech recognisers. Bandwidth economy can be achieved by recogniaing and transmitting these linguistic units.

The difficulties of automatic recognition are discussed and its processes compared with the human mechanism for speech recognition. It is suggested that, just as in human speech recognition, the performance of an automatic recogniser could be improved by using information about the statistics and the structure of the language as well as the usual acoustic cues. The design and construction of a phoneme recogniser for putting this idea to the test is described. The machine has three parts: (1) the acoustic recogniser for detecting aome simple phonemic cues, (2) stored knowledge about the digram frequencies of these phonemes, and (3) a device for aelecting the phoneme that is most likely to occur in the light of both acoustic information and of the relevant digram frequencies. The selection is indicated on a typewriter.

The recogniser has a repertory of 13 phonemes: /a:, i:, u:, ə:, t, k, s, ʃ, f, z, m, n, l/, and deals with about 200 English words, spoken in isolation and containing only these phonemes.

The performance was tested by comparing the phonemes at the input with the output, and also by presenting this output to subjects in visual and in acoustic form. It was found that the score for correctly recognised worda increased from 28 per cent to 43 per cent when information about digram frequencies was added to the acoustic cues. The implications of visual and acoustic presentation of the output were examined and the effect of increasing the contextual information at the disposal of the subjects was teated experimentally. Possible future developments in this field of research are reviewed.

TABLE OF CONTENTS

Chapter I

Page

INTRODUCTION: THE BANDWIDTH COMPRESSION PROBLEM 9

Chapter II

SPEECH TRANSMISSION SYSTEMS USING ARTICULATORY CONSTRAINTS: VOCODERS 13

SPEECH TRANSMISSION SYSTEMS USING LINGUISTIC PRINCIPLES TO ACHIEVE BANDWIDTH ECONOMY 20

Chapter III

THE THEORETICAL BASIS OF THE AUTOMATIC SPEECH RECOGNISER TO BE CONSTRUCTED 23

Chapter IV

THE DESIGN AND CONSTRUCTION OF THE AUTOMATIC PHONEME RECOGNISER 28

THE AUTOMATIC RECOGNITION OF THE VOWELS 33

THE AUTOMATIC RECOGNITION OF THE CONSONANTS /m, n, l, s, ʃ/. 40

THE AUTOMATIC RECOGNITION OF THE CONSONANTS /t, k/. 41

THE AUTOMATIC RECOGNITION OF SPACE BETWEEN WORDS 43

THE AUTOMATIC RECOGNITION OF /f/ AND /z/ 44

THE STORAGE AND USE OF LINGUISTIC INFORMATION 46

THE OPERATION OF THE TYPEWRITER AND THE TYPEWRITER MEMORY 51

THE POWER PACKS 57

Chapter V

	Page
THE SPEECH MATERIAL USED FOR TESTING THE RECOGNISER	62
THE TESTING OF THE RECOGNISER	68
THE PERFORMANCE OF THE RECOGNISER: INPUT/OUTPUT COMPARISONS	73
THE EFFECT OF USING MORE THAN ONE SPEAKER	77
THE PERFORMANCE OF THE RECOGNISER: VISUAL AND ACOUSTIC TESTS	79
THE INFLUENCE OF CONTEXT ON SUBJECTS' ABILITY TO INTERPRET THE OUTPUT OF THE RECOGNISER	85

Chapter VI

CONCLUSIONS	87
REFERENCES	90

LIST OF TABLES

		Page
Table 1.	Voltage and current requirements of the recogniser circuits.	58
Table 2.	Word list 1. (139 words).	62
Table 3.	Frequency of occurrence of phonemes in List 1.	64
Table 4.	(a) Digram frequencies of phonemes in List 1.	65
	(b) Voltage settings of potentiometers representing the digram frequencies in Table 4 (a).	66
Table 5.	Word List 2. (75 words).	67
Table 6.	Word List 3. (200 words).	69
Table 7.	Frequency of occurrence of phonemes in Lists 2 and 3.	71
Table 8.	Confusion matrices for input/output comparisons, using List 1.	74
	(a) Unbiased operation.	
	(b) Biased operation.	
Table 9.	Analysis of errors made by the recogniser, using List 2.	75
Table 10.	Confusion matrix for input/output comparison, using List 3.	78
Table 11.	Comparison of results obtained when using three different speakers.	79

		Page
Table 12.	Key to the transliteration used in the visual presentation of the recogniser output.	81
Table 13.	Comparison of results obtained by the various methods used for testing the performance of the recogniser.	82
Table 14.	Confusion matrices obtained from the responses to visual and acoustic presentations of the biased recogniser output.	83
Table 15.	List of words whose meanings have something in common.	86

LIST OF FIGURES

		Page
Fig. 1.	Principle of operation of Dudley's channel vocoder.	13
Fig. 2.	Principle of operation of baseband channel vocoder.	15
Fig. 3.	Principle of operation of resonance vocoder.	18
Fig. 4.	Principle of operation of baseband resonance vocoder.	19
Fig. 5.	Block diagram illustrating the principle of operation of automatic phoneme recogniser.	28
Fig. 6.	Circuit diagram of filter-bank amplifier. On this and all other circuit diagrams the values of resistance are shown in ohms and of capacitance in microfarads, unless otherwise stated.	30
Fig. 7.	Circuit diagram of analysing filters.	30
Fig. 8.	Frequency response curve of three typical filters and serial numbers with centre frequencies of all 18 filters.	31
Fig. 9.	Circuit diagram of filter rectifier and smoothing filter.	32
Fig. 10.	Pen recording of rectified filter outputs for a few test words.	33
Fig. 11.	Schematic diagram of spectral pattern matching device.	34
Fig. 12.	Circuit diagram of triangular wave generator.	35

		Page
Fig. 13.	Oacillograms demonatrating the operation of the multiplier. (a) Waveshape of triangular voltage generator. (b) Composite oacillogram showing triangular voltage and a square slice from it. (c) and (d) Examples of mark-apace ratioa obtained when slicing at different levels. (e) A aquare wave with the same mark-space ratio as in (c) but limited to a different amplitude.	36
Fig. 14.	Principle of operation of multiplier.	37
Fig. 15.	Circuit diagram of typical multiplier section.	37
Fig. 16.	Maximum detector circuit. (a) Simplified diagram. (b) Detailed circuit diagram of one of 16 identical aections.	39
Fig. 17.	Simplified circuit diagram of the /t/ detector.	42
Fig. 18.	Simplified circuit diagram of the space detector.	43
Fig. 19.	Pen recording illustrating the action of the space detector. (a) Speech envelope. (b) Space detector output. (c) 20 c/s time marker.	44
Fig. 20.	Diagram of amplitude discriminating circuit for /f/ detector.	45
Fig. 21.	A typical section of the "store of linguistic knowledge" circuit.	47
Fig. 22.	Circuit diagram of a typical phoneme memory unit.	48

		Page
Fig. 23.	Schematic diagram showing the arrangement by which acoustic and linguistic information are combined in the recogniser.	50
Fig. 24.	Example of coding network. The 50 volt supply is obtained from the "typewriter output" line of the phoneme memory (Fig.22.) and the single switch on the diagram represents the combination of the contacts of relays A (off) and B (on) of Fig.22.	52
Fig. 25.	Simplified circuit diagram of "write" and "read" uniselector connections.	53
Fig. 26.	Simplified diagram of circuit for operating typewriter solenoids from binary coded control input voltage.	54
Fig. 27.	Circuit diagram of thyratron unit for energising the typewriter solenoids.	56
Fig. 28.	Simplified circuit diagram of 300 volt series stabiliser.	57
Fig. 29.	Simplified circuit diagram of series stabiliser using transistor and valve.	59
Fig. 30.	Circuit diagram of all-transistor series stabiliser.	60
Fig. 31.	The automatic phoneme recogniser.	61
Fig. 32.	Typical recogniser outputs.	72

Acknowledgment

Figures 1, 2, 3 and 4 appear in Recent Developments in Acoustics (Elsevier Publishing Co., Amsterdam, 1961), edited by the late E.G. Richardson and E. Meyer, and are reproduced by courtesy of the publishers.

CHAPTER I

INTRODUCTION: THE BANDWIDTH COMPRESSION PROBLEM

In conventional telephony the sound wave produced by the speaker, after transformation into electrical changes, is transmitted along the line to the receiver where the electrical wave is re-converted into a sound wave similar to that produced by the speaker. The spectrum of speech waves extends over a band roughly 10,000 c.p.s. wide (21) and the intensity variations can be as high as 40 db.* The transmission of high quality speech by conventional means requires a telephone line of the above bandwidth and signal-to-noise ratio, although in normal telephony a 3500 c.p.s. band and a dynamic range of about 30 db. has been found sufficient for the transmission of highly intelligible, if not very natural-sounding, speech. The information-carrying capacity of such lines is high. Shannon's (48) formula gives the channel capacity as

$$C = W \log (1 + P/N)$$

where W = bandwidth of line and P/N = signal-to-noise ratio of line.

This represents about 133,000 bits per second capacity for a line suitable for high quality speech transmission and 35,000 bits per second for "telephone" quality transmission. Quite simple considerations, given below, show that, at least theoretically, a considerable reduction in channel capacity requirements should be possible without influencing the intelligibility of the transmitted speech. If the channel capacity required by one conversation could be reduced, then by using a suitable coding procedure several conversations could be transmitted simultaneously over the same line that previously carried only one conversation. The possible transmission economies resulting from such a system are the reason for the interest in finding ways of reducing the channel capacity required for speech transmission. Systems making use of such principles are called bandwidth economy speech transmission systems and since, as will be seen later, the necessary processes involve the selective transmission of only certain characteristics of the original wave and the reconstruction of a sound wave from the transmitted properties, such systems are also referred to as analysis-synthesis telephone systems.

It has been stated above that telephone lines with a 10,000 c.p.s. bandwidth and a 40 db. dynamic range are being used for the transmission of high quality speech. Such lines have the high information-carrying capacity already mentioned because they are capable of transmitting distinguishably any one of the very large number of different** sound waves that fall within this frequency and intensity range. The variety of such sound waves is very much greater than the number relevant for speech transmission as there is a definite upper limit to the variety of sounds that the human vocal organs can produce. The variety of possible speech sound waves is limited by the well-known fact that their generation is controlled by the movement of a small number of organs, the lips, tongue, teeth, soft palate and vocal cords, all of which are capable of relatively slow movements only. As the number of different

---

* H. Fletcher (21) states that the range of intensities of speech sounds in normal conversational speech is about 30 db. and an additional over-all intensity variation of 10 db. then makes the dynamic range of normal speech about 40 db.

** "Different" means the distinguishable steps permitted by the noise level.

speech sound waves that can be generated, and therefore can be available for transmission along the telephone line, is smaller than the number of different sound waves contained within the 10,000 c.p.s. bandwidth, it should be possible to transmit speech, without the loss of intelligibility, along a line with a smaller channel capacity. This must be so because the requirements for channel capacity are a function of the total number in the ensemble of messages from which the transmitted signal is selected. A further reduction in the theoretical requirements for channel capacity results from the fact that the vocal organs change only gradually: as a result, successive speech wave patterns are often highly correlated and such a set of signals represents a smaller information rate than if the various possible messages followed each other randomly. Further evidence for assuming that the transmission of speech does not require the full channel capacity of a line with a 10,000 c.p.s. bandwidth is given by Gabor (26) when he points out that the human hearing mechanism could probably not assimilate information received at the high rate at which a telephone line is capable of transmitting it.

The channel capacity required for the transmission of speech-like sound patterns can be estimated by assuming that the information-carrying characteristics of speech waves are determined by the movement of five independent vocal organs (vocal cords, lips, tongue, teeth, soft palate), that information about their position is to be transmitted with a 3% accuracy (or 30 db. signal-to-noise ratio) and that their maximum rate of movement represents 50 significant changes per second. The latter figure assumes that on the average as many as five positions of the vocal organs need be known for defining each phoneme, with the phonemes produced at the relatively high rate of 10 per second. For these numerical values the informational content of each vocal organ position is

$$\log_2 1000 \doteq 10 \text{ bits}$$

and the channel capacity required for the transmission of information about the movement of each vocal organ is

$$50 \times 10 = 500 \text{ bits per second.}$$

This gives a total channel capacity requirement of 2,500 bits per second for the transmission of all significant articulatory changes in speech, with no loss of intelligibility. The above figures probably under-estimate the requirements of the vocal cord channel, but over-estimate those relating to the other vocal organs. The channel capacity requirements would be greater if information about individual voice characteristics were to be transmitted as well. The transmission of such information would probably require mainly a closer specification of the action of the vocal cords and would not necessarily increase the channel capacity requirements to any large extent. The above estimate of 2,500 bits per second is a considerable reduction compared with the 133,000 bits per second for high quality transmission and the 35,000 bits per second for telephone quality lines. These theoretical possibilities of bandwidth economy can be realised only if the telephone line is restricted to the transmission of speech-like sounds and such a line would not be able to deal with any other type of sound, such as music for example, or at least not without severe distortion. The realisation of such a telephone system requires the automatic recognition of the acoustic correlates of all distinguishable articulatory changes - that is to say of the significant characteristics of the speech sound wave - and their conversion into a code suitable for the full exploitation of the channel capacity of the line; at the receiving end the code would be re-converted into a speech sound wave. Neither the relevant acoustic characteristics nor methods for their automatic extraction are fully known yet, but work in this direction is being actively pursued with a view to making possible this type of bandwidth economy speech transmission.

Further bandwidth economies should theoretically be possible because of the linguistic origin of the speech sound waves. When communicating, the speaker first organises the message to be transmitted in linguistic form, in terms of the phoneme sequence and also stress, intonation, etc. These linguistic units are translated successively into other forms and eventually into a speech sound wave which is the acoustic form of the original linguistic code. The sound wave, on reaching the listener, stimulates his hearing mechanism and the acoustic code of the transmitted message is converted into a neural one. This is eventually re-converted into the linguistic code in the listener's brain. It is this linguistic code which is then interpreted as meaningful information. During the initial and final stages of this transmission sequence the message is encoded in phonemic form and no greater channel capacity should be necessary for the transmission of speech than is needed for the transmission of the phonemes. Taking the number of phonemes in English as 40 and assuming that they occur at the rate of 10 per second then the channel capacity required for transmitting the phonemic information is

$$10 \log_2 40 = 53 \text{ bits per second.}$$

The phonemic sequence does not, however, contain as much speech information as the speech sound wave. Information about the identity of the speaker, his emotional state, his attitude to the subject-matter, such things as emphasis, doubt, assertion, etc. are transmitted by the speech sound wave, but not on the whole by the phonemic sequence. It is possible to make a very approximate estimate of the channel capacity required for the transmission of these features of speech, again based entirely on theoretical views of the structure of English, the variety and speed of variation of the units used by the language structure for formulating these aspects of information, rather than on our present state of knowledge about which of these units can in fact be recognised, automatically or otherwise. Intonation, for example, probably uses less than 10 units that follow each other not faster than 2 or 3 times per second and therefore a channel capacity of 10 bits per second should be sufficient for its transmission. After making similar estimates for stress and rhythm, it seems reasonable to assume that a channel capacity of not more than 100 to 150 bits per second should be sufficient to transmit information about the sequence of the various structural units of English as they occur in normal speech. This is considerably less than the 2500 bits per second estimated for transmitting information about the articulatory movements of speech or the 35,000 to 133,000 bits per second for transmitting the speech sound wave *in toto*.

It may be of interest to interrupt the main argument at this stage to point out that the above calculations give information rates for speech that are still higher than the theoretical minimum, because they assume that successive speech units can occur in random order. In fact, owing to a variety of linguistic rules, extensive statistical laws affect the possible sequences of units. These sequential probabilities can be exploited, at least theoretically, for a further reduction of the channel capacity requirements for speech transmission. Similar arguments apply to the transmission of articulatory information, because on the whole the sequence of articulatory configurations is also statistically determined: certain sequences do not occur at all and others again have differing probabilities of occurrence. These statistical relationships of successive articulations can be exploited to reduce information rates. One way of doing this has been proposed by C. P. Smith (49).

Returning now to the discussion of the linguistic organisation of speech information, it seems that this information, when expressed in terms of linguistic units, requires a much smaller channel capacity for transmission than at the articulatory or the acoustic levels. If the linguistic nature of speech is to be utilised in order to

achieve bandwidth economy, then the terminal equipment at the sending end must sort the input waves into categories corresponding to linguistic units, rather than to articulatory configurations. The information about the linguiatic units must be coded efficiently and transmitted; at the receiving end a synthesiser is needed which is capable of generating a aound wave which will make a listener recognise the same linguistic unit that the code applied to its input represented. An idea of the nature of such a system can be gained by considering what happens when a speaker dictates to a teleprinter operator. The teleprinter operator receives the aound waves produced by the speaker and transforms the speech information from the acoustic into its linguistic form - in other worda, he understands what is being said to him. The linguistic units, for instance in the form of a sequence of phonemes or letters, are then typed on the teleprinter by the operator, producing an electrical code with a one-to-one correlation with the letters being sent. At the receiving end another person could read back aloud the message being typed out, thereby re-converting the linguistic (visual) data into its acoustic form. The bandwidth required for such a transmission system is relatively small, corresponding to the amall variety and slow rate of change of the units, the letters, being tranamitted. The all-important transformation of information from the acouatic to the linguistic form and back again ia in this case carried out by a human being at either end of the transmission channel. The human being has no difficulty in carrying out these transformations, although he is not able to formulate the rules he uses for this process. Machines that perform the same operations as the above human being are called automatic speech recognisers and speech synthesisera reapectively. The rules for constructing satisfactory devices of thia kind are not known yet and the search for these rules is an important part of experimental phonetica and finding them is an essential pre-requiaite for the design of this type of bandwidth economy telephone system.

In the foregoing discuasion it has been pointed out that two basically different types of constraint, articulatory and linguistic, can be utilised in the design of speech transmission syatems requiring less bandwidth than that needed for transmitting the speech sound wave itself. In each case terminal equipment is needed which sorts the input waves into a number of classes that are smaller than the total number of different waves possible within the speech spectrum. The principle of this sorting process is basically different in the two methods. One attempts to derive a simplified description of the speech sound wave that is based on the operation of the articulatory mechanism and therefore only knowledge of articulatory-acoustic correlations is needed and no linguistic considerations whatsoever are involved. The other system is concerned solely with a description of the speech information in terma of that sequence of linguistic units on which the production of the sound wave was baaed. The systems using the first of theae two principles are called vocodera. They have been extensively investigated in the past and a short deacription of their operation can be found in the next chapter. This is followed by information about automatic speech recognisers and about efforts for using them in speech transmission systems of reduced bandwidth - as far as they have been described in the literature. These latter systems are neither numerous nor very successful: unfortunately the characteristics of the sound patterns associated with the various linguistic units are variable and overlapping and it is not easy therefore to recognise them by machine. The question of how far a particular principle, associated with the linguistic nature of the speech input, can be used to improve the performance of automatic speech recognition is discussed next; the construction of an automatic recogniser incorporating this principle is described, as well as the results obtained in experiments for testing its performance and finally poasible future developments in this field of research are reviewed. Such work should throw some light on the operation of the human speech recognition process and at the same time help in the design of bandwidth compression telephone systems. Once such a recogniser is achieved it can be used as a terminal converter at the sending end of a

speech transmission system in which the electrical signals transmitted correspond to the phoneme sequence.

CHAPTER II

SPEECH TRANSMISSION SYSTEMS USING ARTICULATORY CONSTRAINTS: VOCODERS

All vocoder systems, in their design, take into consideration the basic mechanism of speech sound production, as put forward by Homer Dudley (10). In speech sound generation the energy of the air flow from the lungs is converted into audible, alternating sound pressure by the action of one mechanism or another to be called the "sound source". Principally one of two different sound sources is active. One of these is provided by the action of the larynx which produces a train of pulses where the pulse repetition rate and to some extent also the shape of the pulses is variable. The other sound source is the hiss produced when the stream of air from the lungs is forced through a narrowing of the vocal tract and also when a stream of fast-flowing air hits an obstacle like the teeth. The spectrum of the hissy sound is random in character (31). These sound sources may be active on their own or in combination. The spectrum of the sound waves generated by these sound sources is modified by the acoustic impedance of the vocal tract which in turn is determined by the articulatory configuration, that is the position of lips, teeth, tongue and palate. These articulatory organs can move only relatively slowly and therefore the spectral changes take place at a correspondingly slow rate.

The first of the vocoders, the so-called channel vocoder developed by Dudley (9) (10), utilises some of these features to obtain a more economical description of the speech wave. The basic principle of its operation will be seen from Fig. 1. At the

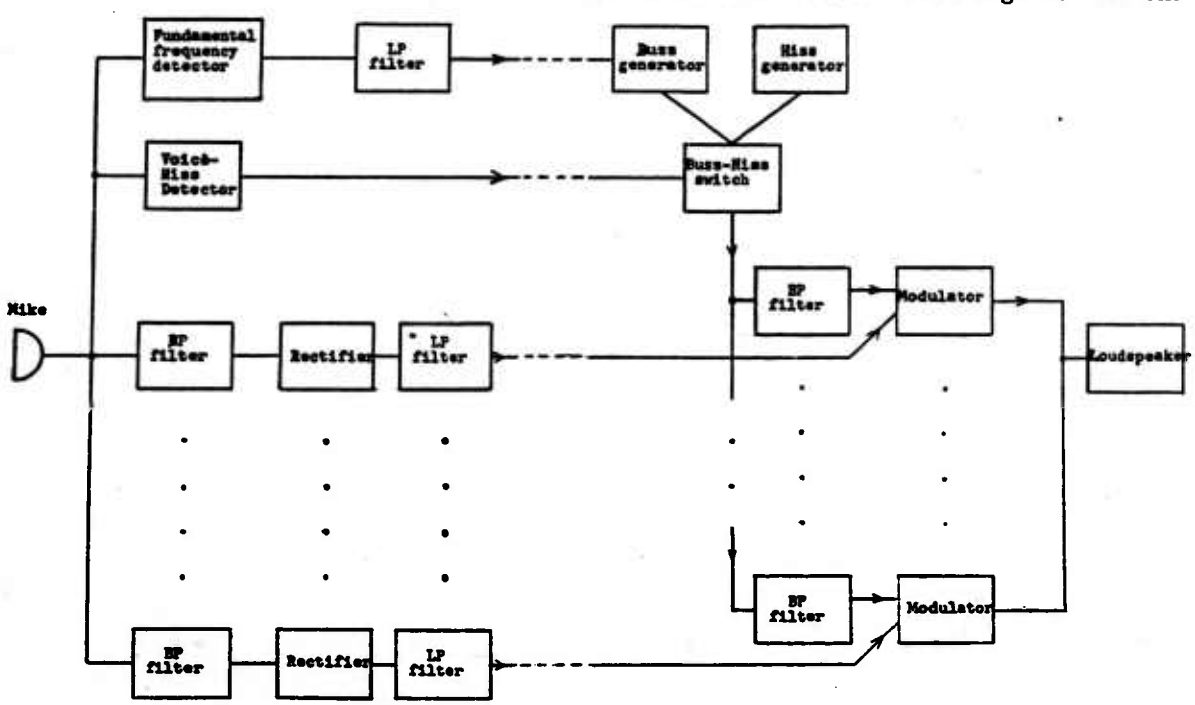


Fig. 1. Principle of operation of Dudley's channel vocoder.

sending end there is a "voice-hiss discriminator" which determines whether the sound wave has been produced by the larynx or the hiss source. The bulk of the energy produced by the larynx source is in the low frequency region of the acoustic spectrum, whilst the hiss sounds are on the whole concentrated in the higher frequency regions and the operation of the "voice-hiss discriminator" relies on this spectral cue. Whenever laryngeal excitation is indicated the "fundamental frequency detector" provides an additional voltage which is proportional to the pulse repetition (larynx) frequency. Dudley used the conventional zero-crossing count to provide an indication of this frequency. Information about the spectral envelope, and therefore about the position of the articulators, is provided by the 10 channel filters. These divide the 100 to 3000 c.p.s. band into adjacent sections and the rectified, smoothed output of the filters indicates the level of energy falling within that section of the spectrum. Since the articulators move only slowly, the spectrum will change correspondingly slowly; hence the filter-rectifiers can be followed by smoothing filters with a low cut-off frequency and the outputs will still indicate all significant changes in the spectrum. Information about the source function, the fundamental frequency and about the output of each filter is transmitted along separate channels, each 25 c.p.s. wide. At the receiving end the information is used to synthesise a sound wave similar to that at the input. A pulse generator and a white noise generator are available as alternative sound sources. A switch, controlled by the signal from the "voice-hiss discriminator" at the sending end, connects one or other of these source voltages to a bank of 10 filters which are similar to the 10 analysing filters at the sending end. The frequency of the pulse generator is made the same as the fundamental frequency of the speech wave being transmitted by making use of the information transmitted from the fundamental frequency detector. The synthesising filters divide the energy of the sound source into 10 spectral bands. Information transmitted along the remaining channels is used to control the output level of each synthesising filter so that it is the same as that of the corresponding analysing filter. Finally the outputs of all the filters are added and applied to a loudspeaker. Dudley's vocoder required a total bandwidth of about 300 c.p.s., offering a 10 : 1 bandwidth compression compared with the normal telephone channel. This economy in bandwidth is achieved by allowing only certain kinds of source function, by transmitting only slow variations of the spectral envelope and by transmitting spectral data averaged over a limited number of frequency intervals, that is, by making use of constraints arising out of the nature of the vocal mechanism. In addition, only details of the amplitude spectrum but not of the phase spectrum are transmitted; in this way a constraint resulting from the nature of the hearing mechanism is also exploited (since the ear does not make use of phase information).

On testing the speech transmission efficiency of the channel vocoder Dudley found that a word articulation of about 70% could be achieved. This already very good performance was later improved as a result of extensive research carried out at the Bell Telephone Laboratories and at the British Post Office (4) (29) (54). The latest vocoders use 18 channels, with cut-off frequencies of about 20 c.p.s., requiring a total bandwidth of about 350 c.p.s., a 30 db. signal to noise ratio and show a 90% articulation score when tested with PB words.*

The naturalness of the transmitted speech, as distinct from its intelligibility, was not very good. Naturalness is a sensory dimension which was at that time, and for that matter still is, not well-defined and no generally accepted methods for

---

* Phonetically balanced word lists deeigned by the Psycho-acoustic Laboratories, Harvard University and published by J. Egan, Laryngoscope, *58*, 955-991, 1948.

measuring it were, or are, available. The naturalness of the system was not considered as good as that of conventional telephone systems and largely for this reason the vocoder is still not in commercial use. A considerable amount of work has, of course, been done to improve the naturalness of the original channel vocoder. Experimental evidence suggests that two of the major causes of this lack of naturalness are incorrect switching of the buzz and hiss source generators and failure of the fundamental frequency of the output to follow the varistions of the larynx vibration frequency of the speaker. In other words lack of definition of the excitation function rather than insufficient information about spectral envelope is at fault and the remedy lies not so much with a different design of channel filter as with an improvement of the buzz-hiss switching and fundamental frequency detector circuits. Several attempts have been made (6) (28) to improve the performance of the fundamental frequency detector circuits without producing a really satisfactory solution. A number of factors make the measurement of the fundamental frequency of the speech wave difficult. One of these is the presence of strong harmonics, another is that the averaging process, which is part of most frequency meter circuits, smooths out the cycle-by-cycle variations of fundamental frequency that can occur in speech and that are sometimes significant. Recently an autocorrelation method has been tried (27) to overcome these difficulties. The delay which produces the greatest value of autocorrelation indicates the fundamental frequency and the actual value of the autocorrelation for this delay is used to control the buzz-hiss switch.

The so-called base band (47) vocoder represents yet another method for obtaining the correct excitation function at the receiving end of the vocoder. As can be seen from the schematic diagram in Fig. 2, a narrow section of the spectrum of the original

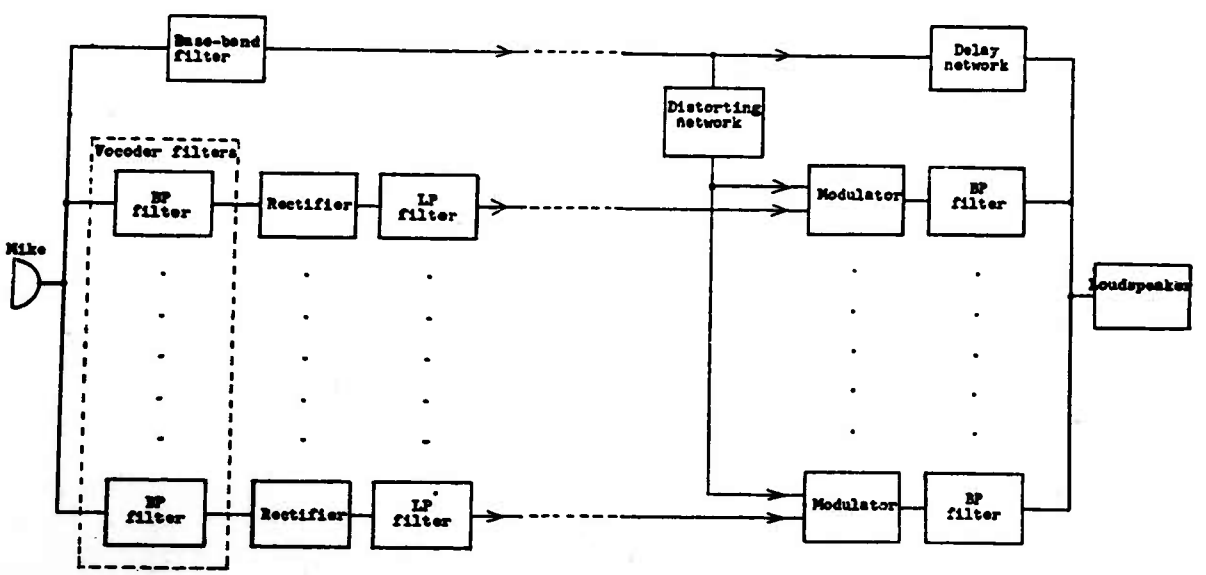


Fig. 2. Principle of operation of baseband channel vocoder.

speech wave is transmitted separately. At the receiving end, suitable circuitry converts this small part of the original speech spectrum into a wave having a uniform spectrum extending over the whole speech frequency range, but at the same time maintaining the periodicity or hissy character of the original speech wave. This converted wave is used to excite the channel filters. In this way the synthesised output will always have the same fundamental frequency as the original speech input and no buzz-hiss switch is required because the excitation function is part of the original speech wave. Probably a 400 to 500 c.p.s. wide slice of the original speech spectrum has to be transmitted for obtaining satisfactory results, and therefore a bandwidth of about 650 c.p.s. is required for the vocoder as a whole. This seems rather a lot of bandwidth for transmitting details of the excitation function of speech when compared with the 300 c.p.s. bandwidth required to transmit information about the spectral envelope. In fact it is not much more than the theoretical minimum because it seems that one of the factors that affect the naturalness of speech is the correct reproduction of the sometimes quite large cycle-by-cycle changes in the fundamental frequency of speech.

The channel vocoder exploits only some of the constraints that arise out of the nature of the speech-producing mechanism: the fact that only certain kinds of excitation function are possible and that the spectral envelope can vary only at a relatively slow rate. This will reduce considerably the number of different sound patterns for transmission and therefore the channel capacity required, but the number is still much larger than the variety that the vocal organs can actually generate. A good illustration of this fact is given by David (4). He points out that the output of a 16-channel vocoder in which the amplitudes in each channel are quantised to eight discrete levels can represent $8^{16} = 2^{48}$ different spectral envelopes; if the vocoder produced as many as 32 different patterns per second it would still take more than 10^5 years to produce all possible patterns. An estimate of the variety of sounds that can be produced by the vocal organs and, at the same time, a guide for a method of classifying sound waves into categories that represent all possible speech waves but no others is provided by the results of research on the acoustics of the vocal organs. Numerous experiments and computations (15) (38) have shown that, in the majority of cases at any rate and in particular for all vowel-like sounds, the spectral envelope of the acoustic output of the vocal organs can be specified in terms of the frequencies of the first three peaks of the spectrum. These spectral peaks correspond to the resonances of the vocal tract impedance and are called the formants. It has also been shown (14) that, as long as the formant frequencies are specified, no further information is required for determining the relative amplitudes of these formants. It is also interesting to note at this stage that just as on the acoustic level the spectral envelope can be specified by stating the values of only three separate variables, the three formant frequencies, so on the articulatory level the configuration of the vocal tract can be specified by defining three variables, namely the distance from the glottis to the greatest constriction along the vocal tract, the size of this constriction and the configuration of the lip opening (13) (14) (51). Once it has been established that the spectral envelope of speech waves can be specified just by the values of the three formant frequencies, then the possible number of significantly different spectral patterns can be calculated as long as it is known to what accuracy the formant frequencies need be known. A good guide to the accuracy required is the ability of the listener to detect changes of formant frequency and this has been determined by experiments in which speechlike sounds with a variety of values of formant frequency were produced using a "terminal analogue" synthesiser (16). It was found that the threshold of discrimination was about ±3%. This would mean that only 33^3 or approximately 2^{15} different spectral envelopes can be distinguished when dealing with vowel-like sounds as compared with the variety of 2^{48} patterns that can be specified by the output of a channel vocoder. Specification of the spectral envelope, of course, does not provide enough information for reconstruction of the sound wave itself: the

source (excitation) function and the absolute amplitude (ss opposed to the relative amplitudes of the pesks) hsve to be known slso.

A speech transmission system then suggests itself, in which, at the sending end, terminal equipment measures the frequencies of the first three spectral peaks; it also determines the characteristics of the source function, whether it is periodic or hissy and the value of the fundamental frequency, if any, as well as the overall amplitude. The results of these measurements should be sufficient to specify the speech sound wave to be transmitted. The relevant information csn be transmitted along only five channels, three for the three formant frequencies, one for the overall amplitude and one for the fundamental frequency which indicates st the same time whether the source function is periodic or hissy. The signal-to-noise rstio required for each channel is indicated by the ability of the human listener to distinguish changes in the speech sound wsve. This ability has been determined in a series of experiments (16) (17) (19) and the vslues found range from ±2% for fundamental frequency through ±3% for formant frequency to ±12% for amplitude and indicate that a signal-to-noise ratio of about 30 db. should be sufficient for all but the fundamental frequency channel which needs a 35 db. dynamic range. The spectrsl configuration can still change only slowly so that each transmission chsnnel needs no more thsn the 20 c.p.s. bandwidth specified for the channel vocoder. At the receiving end, a sound wave similar to the original speech wave can be genersted using s suitable synthesiser controlled by the information trsnsmitted along the five channels. The synthesiser is usually a "terminal analogue" of the humsn vocsl trsct. There is a pulse generator and a hiss generstor to provide alternstive source functions, and they are connected to the spectrum-forming resonant circuits by the buzz-hiss switch. Just as in the chsnnel vocoder the vslue of the pulse frequency snd the stste of the buzz-hiss switch are controlled by information transmitted from the sending end. The resonant frequencies of the tuned circuits sre variable and each is varied, using the information transmitted from the sending end, to correspond with one of the three formant frequencies of the speech input. In this way the pesks of the spectrum of the output will always be at the same frequencies as the formants of the speech input. The three resonant circuits can be connected in series or in parsllel. If they are connected in series, then the amplitudes of the formant peaks are automatically sdjusted as the energy goes from one resonant circuit to the other (14), snd informstion about oversll smplitude only is required: in the parsllel connection the resonant circuits do not affect each other and therefore separate information is required about the amplitude of each formsnt. Therefore fewer transmission chsnnels are required with the series connection; st the same time sn error in transmitting the vslue of a formant frequency will disturb a series system more than a psrsllel one becsuse in the former system the transmitted values of the formant frequencies also control the amplitudes of the formant peaks in the output.

A schematic diagrsm showing the principsl constituents of such speech transmission systems, usually called resonsnce or formant vocoders, is shown in Fig. 3. Several resonance vocoder systems have been tried (1) (18) (32) (50); they sll use the same basic principles although the circuitry which implements these principles varies considerably. At the sending end the electronics for extrscting information about the source function is similar to that used in the chsnnel vocoder. The formant frequencies are determined by first dividing the spectrum into three bands in which the three formants sre expected. The value of the spectrsl peak in each of these three bands is then found either by sub-dividing by further filtering and finding the filter with the maximum output or by a method based on finding the average value of the rate of zero-crossings. At the receiving end Miller capscitiea,

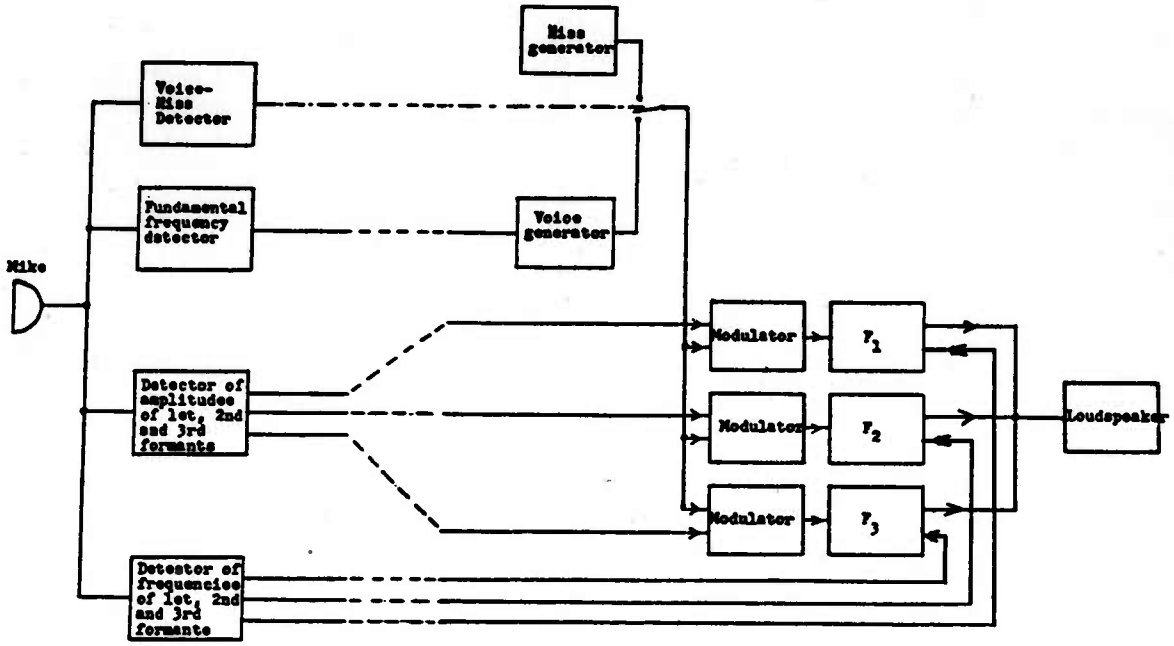


Fig. 3. Principle of operation of resonance vocoder.

variable inductances (increductors) and many other devices have been used to obtain variable tuning for the resonant circuits. A typical resonance vocoder circuit gives 70% to 80% articulation with PB words and needs a bandwidth of 100 to 150 c.p.s.

Although these figures show that resonance vocoders transmit speech with good intelligibility they are afflicted with the same lack of naturalness as the channel vocoder. This is not surprising as they use the same circuits for dealing with the excitation function as the channel vocoder. Consequently the base band principle, which gave good results with the channel vocoder, has also been tried with the resonance vocoder (20). The schematic diagram of such a device is shown in Fig. 4. It requires a bandwidth of about 550 c.p.s. and combines 80% intelligibility for PB words with naturalness that is noticeably better than that of the conventional channel vocoder and at the same time still provides a bandwidth compression of about 5 or 6 to 1.

It may be of interest to diverge at this point and explain that although the resonance vocoder seems to be able to transmit most speech sounds with a reasonable degree of intelligibility, it is still primarily a system suitable for the transmission of vowel-like sounds. This is because at the sending end the sound waves to be transmitted are categorised in terms of the first three spectral peaks and similarly at the receiving end a terminal analogue synthesiser is used whose output is characterised by the corresponding three spectral maxima. It is only the vowel-like sounds, in other words sounds generated with the sound source at the far end of the vocal tract and with the vocal tract consisting of a single tube without side branches, that can be described in terms of spectral maxima only (15). As soon as the sound source moves further forward or when side branches of the vocal tract come into play, as is the case for the generation of fricatives and nasals, the

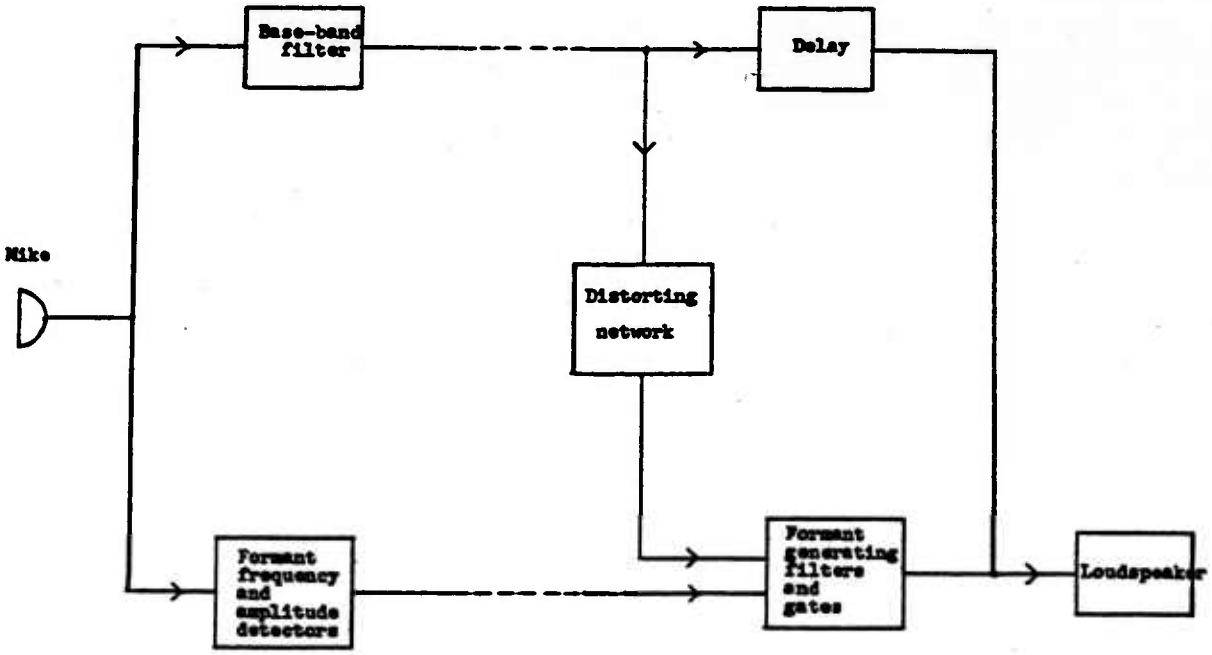


Fig. 4. Principle of operation of baseband resonance vocoder.

spectrum of the sound wave generated is characterised by minima as well as by maxima. No adequate methods of analysis are available as yet which would furnish the characteristics of these spectra. Also, the terminal analogue synthesiser in the form in which it is being used in existing resonance vocoders could not generate the corresponding sounds and a vocal tract analogue type synthesiser would be needed.

If a vocal tract analogue is to be used then it may be more convenient to control its adjustment in articulatory terms rather than acoustic ones, that is by specifying the tongue and lip positions, etc., to which it is to be adjusted rather than the spectrum of the resulting sound wave. This is only possible if the acoustic information about the speech sound wave can first be interpreted in articulatory terms and vice versa. Published information for performing such a conversion is as yet available for certain classes of sound only (51) (52) and transmission systems of this kind have not been tried.

One more vocoder has to be described, the pattern vocoder (49), which is only now being developed. It attempts to reduce the channel capacity required for speech transmission by ensuring that only those spectral patterns that actually do occur in speech can be transmitted: those spectral patterns that in preliminary experiments have been found to contribute to speech intelligibility and are stored for use in the transmission system. The speech input is first of all applied to a bank of filters, just as in the channel vocoder. The spectral envelope represented by the output of these filters is sampled 50 times per second and the samples are then compared with each of the stored patterns. The identity of whichever stored pattern matches best with the spectrum of the input is indicated and a corresponding serial number is transmitted. At the receiving end this serial number is used to select a set of stored control instructions which adjust a synthesiser so as to produce a

sound wave with the corresponding spectral envelope. This process is repeated for each sample of the input spectrum. The analyser at the sending end also provides the usual information about the sound source, buzz-hiss distinction, value of the fundamental frequency and amplitude and this information is transmitted and used at the receiving end to control the operation of the synthesiser. It is hoped that not more than 4,000 different spectral patterns will have to be stored to make possible the transmission of intelligible speech and that the input spectrum would have to be sampled not more than about 50 times per second; this represents an information rate of 600 bits per second. The system as described so far can transmit any sequence of these stored patterns and does not exploit the fact that in speech these spectral patterns do not occur in random sequences. Further bandwidth economy may be possible if in addition to the store of spectral patterns already described the *sequences* of spectral patterns that have been observed in speech are also stored. The patterns recognised are not immediately transmitted but remembered and the observed pattern-sequences are compared with the stored sequences; the identity of the stored sequence that agrees best with the input sequence is indicated and transmitted. At the receiving end this information is used to select one of a set of stored control instructions which will adjust the synthesiser to reproduce the correct sound sequence at the sending end. Experiments have yet to be carried out to find how many spectral patterns, how many patterns per pattern sequence and how many pattern sequences are needed for satisfactory speech transmission. The saving in bandwidth cannot be assessed until these figures are known, but such a system is bound to be more economical in bandwidth because of the inevitably large number of sequences that do not occur in speech which other types of vocoder are capable of transmitting and the pattern vocoder is not.

The preceding paragraphs summarize the operation of the principal vocoder systems that have been or are being tried. The principle by which they achieve bandwidth economy is shared by them all: using our knowledge of the operation of the vocal organs and of the human perceptive mechanism and without losing any precision in specifying those characteristics that are relevant to intelligibility, the sound input is classified into a smaller number of categories and significant variations are restricted to a slower rate than would be possible for the bandwidth and signal-to-noise ratio of the original speech wave. This simplified description of the input is transmitted and then used to control a synthesiser in such a way that a sound wave similar to that at the input is generated. Our knowledge is not yet extensive enough to obtain a specification of articulatory action from the acoustic wave. Work is proceeding in this direction and results may perhaps be useful in achieving further bandwidth compression.

SPEECH TRANSMISSION SYSTEMS USING LINGUISTIC PRINCIPLES TO ACHIEVE BANDWIDTH ECONOMY

As was pointed out earlier, much greater economies in channel capacity could be achieved if the acoustic input was classified in terms of linguistic categories. As a rule quite a number of acoustic patterns (or of articulatory configurations) or even sequences of these correspond to one linguistic unit. Therefore the total number of possible linguistic units is smaller, or they vary at a lower rate or both, and these linguistic units can be transmitted over lines with smaller channel capacity. Unfortunately the rules for classifying the acoustic patterns into categories corresponding to linguistic units are not yet known and neither are the rules for the reverse process, for controlling a speech synthesiser from a phonemic input. One attempt has recently been made to summarise our knowledge of phonemic synthesis (42) but the construction of a practical synthesiser based on these rules has not

yet been tried. As far as linguistic recognisers are concerned, several attempts have been made to classify speech wave inputs into various linguistic categories such as phonemes or words. Any device which classifies the speech wave input into categories corresponding to linguistic units and then indicates the result by producing a coded signal which has a one-to-one relationship with the linguistic unit is called an automatic speech recogniser; depending on the linguistic unit which forms the basis of this operation there are automatic phoneme recognisers, automatic word recognisers, etc.

The earliest attempt to construct a phoneme recogniser was probably that of Dreyfus-Graf (7) (8) whose method is based entirely on a rough analysis of the spectral envelope of the speech sound wave. The input is applied to band pass filters which divide the spectrum into six adjacent bands and cover the total range of 80 c.p.s. to 3800 c.p.s. The output of these filters, after rectification and smoothing, is used to control the deflection of a pen recorder. The pen can move in any one of six directions, at an angle of 60° to each other and all in the horizontal plane, and the amplitude and direction of the deflection correspond to the resultant of the outputs of the six band pass filters; the pen writes on a sheet of paper that is moved past the pen at a uniform rate. In response to a speech input, the pen will draw figures which should be characteristic of the spectral pattern which gave rise to them. Using one speaker only and for sounds produced in isolation or words spoken slowly, recognisably different patterns were drawn by the pen for different speech sounds, In some versions of his machine, Dreyfus-Graf replaced the pen recorder by a number of differential relays which, depending on the configuration of output of the six filters, operated one of a set of contacts. These in turn operated the keys of a typewriter.

To ensure that a separate figure is drawn for successive phonemes, it is necessary to return the pen to the central position whenever a new phoneme starts. This raises the difficult problem of deciding when one phoneme ends and the next starts: the so-called "gating" problem. "Gating" is a problem common to all phoneme recognisers and arises from the fact that the speech sound wave, a continuous event, has to be analysed in terms of a sequence of linguistic units which are by definition discrete entities, one following the other. Dreyfus-Graf first operated his phoneme-gate after a fixed interval of about 1/20 sec. and this obviously made the shape of the output patterns very dependent on the rate of speaking. Later he used the rate of change of averaged speech intensity variations for gating: a fresh phoneme recognition was started whenever the overall intensity, averaged over a short time interval, showed a fast change. Some versions of Dreyfus-Graf's system used this wave envelope also as additional information for the recognition of consonants. The intensity time function was applied to a bank of filters dividing the 2 c.p.s. to 64 c.p.s. frequency range into four adjacent bands. Dreyfus-Graf called the prominent components of the resulting spectrum "sub-formants" and used them extensively in recognition.

Basically, then, Dreyfus-Graf's method produces a specialised visual representation of the acoustic spectrum and is based on the assumption of a one-to-one relationship between spectral patterns and phonemes. It performs a true recognition, that is, classification, process only in so far as it converts the wave continuum into discrete segments. Within each time segment, however, the visual representation of the acoustic spectrum is along a continuous scale and therefore no classification is in fact carried out by the "recogniser"; instead the classification must be performed by the observer of the written patterns. In this sense therefore this is not really a phoneme recogniser at all but is more like a vocoder. In a vocoder the spectral anslysis is also performed on a con-

tinuous scale, but the process in reversible and the output of the analyser can be used conveniently for the syntheais of a corresponding sound wave which is then presented to a listener; it is the listener who then claasifies these sound patterns into the corresponding phonemic categories by using the normal method of speech recognition at his disposal. Similarly in Dreyfus-Graf's system it is the observer rather than the automatic recogniser (despite the fact that the machine is called a phoneme recogniser) that carries out the classification; the observer finds this a much more difficult task than the classification of acoustic patterns because the framework of normal speech recognition already acquired cannot be used.

On the other hand, true classification of acoustic patterns into groups corresponding to linguistic units is carried out by the automatic phoneme recogniser deaigned by Wiren and Stubbs (55). In their method, the speech input was examined for the presence or absence of acouatic properties thought to be characteristic of certain linguistic classes and the results were used in a succession of binary selections to reach a final classification into phonemic groups. This approach was based on the idea of distinctive features (34). The distinctive features, as proposed by Jakobson, Fant and Halle, are a set of linguistic attributes and the liatener identifies the phonemes by a series of binary decisions based on the presence or absence of some or all of these features. Jakobson, Fant and Halle suggest some acoustic correlates of these linguistic features and Wiren and Stubbs have based the operation of their phcneme recogniser on the detection of these and other acoustic correlates and using them for phoneme identification in a succession of binary deciaions similar to that suggested by the distinctive feature approach. The distinctive oppositions on the linguistic side for which acoustic correlatea were sought, were voiced/unvoiced, stop/fricative, non-turbulent/voiced turbulent, vowel/vowel-like consonant and acute/grave. Most of the acoustic distinctions depended on the spectral distribution of energy but amplitude and rate of rise of amplitude were also used as cues. Different parts of the system were tested with up to several hundred utterances of anything from 4 to 20 different speakers, giving success scores that varied from 50% to almost 100%. The results are a reflection of the fact that the distinctive features and their acoustic correlates are far from being related in a one-to-one manner.

Another system of automatic speech recognition is described in two publications from the Bell Telephone Laboratories (11) (12) in which two versions of basically similar equipment are described. One of them is a spoken digit recogniaer which is concerned with the recognition of the ten numbers 0 to 9, each pronounced in isolation. In the first stage of the recognition process, the speech wave is applied to a bank of 10 filters, similar to those used in Dudley's channel vocoder. In a aeries of preliminary experiments the average spectrum, as represented at the output of the 10 filters, was obtained for 10 different sustained speech sounds, some of them vowels and others consonants, all of them having been found significant in the recognition of the spoken digits. During the recognition process the spectral envelope of the speech input is matched separately with each of the reference patterns obtained from the preliminary experiments. The identity of whichever stored pattern matches the input best is indicated by the operation of the appropriate one of ten relays provided. Whenever during the utterance of the digit the spectrum of the input changes sufficiently for the best match to have shifted to another of the stored patterns, then the previously energised relay releases and the appropriate fresh relay operates. Further preliminary experiments are needed for the last stage of the recognition process: the average duration for which each of the 10 relays is operated during the utterance of each of the 10 digits is determined. The actual duration for which each of the relays is operated whilst the speech input to be recognised is articulated

is compared with each of the duration patterns obtained in the preliminary experiments and the best match is indicated and provides the final choice. This recognition process corresponds in effect to the selection of the best match obtained by comparing the frequency-amplitude-time spectrogram of the input with each of 10 reference spectrograms, one for each of the 10 digits as obtained in the preliminary experiments. The time dimension of these spectrograms takes into consideration only the duration of each spectral element and not the order in which they occur. The machine recognises correctly about 97% of the numbers spoken into it by one speaker and its performance deteriorates when several voices are used. This then is a true automatic word recogniser in which the machine categorises the acoustic input into 10 classes corresponding to linguistic units: words. Also, the machine has a built-in *a priori* knowledge of all the words that are possible in the language it has to deal with, 10 in this case, and recognition is based not on the measurement of some absolute values but on finding that one of the possible categories to which the input is most similar. The high degree of success is, of course, in no small measure due to the fact that only 10 words are possible in the recogniser's language.

A very similar development of this system is an attempt to achieve a phonemic transmission system. Information about the identities of the spectral patterns, recognised in the manner of the digit recogniser just described, are coded and transmitted. At the receiving end this information is used to operate a synthesising circuit which will generate a sound wave with an appropriate spectrum. In this way the channel capacity of the transmission system need be no greater than what is required for the transmission of one of 10 symbols following each other at some slow rate, say 15 per second. The system was first tried with only the 10 spoken digits as before and two listeners recognised almost all the words spoken. When the vocabulary was increased to include another 37 mono-syllabic words, then the score dropped to about 50%. This system, much extended in several ways and using digital methods, is being tried again (49) and seems promising.

CHAPTER III

THE THEORETICAL BASIS OF THE AUTOMATIC SPEECH RECOGNISER TO BE CONSTRUCTED

It is clear from the description of the existing systems that a method for really successful automatic speech recognition has not yet been found. In the search for a solution it has always been realised that phonetic context and other variables will influence the acoustic features that characterise the phonemes and words. It has always been tacitly assumed, however, that there are some invariant acoustic features that characterise a phoneme and that are always present when that particular phoneme is spoken by the speaker or recognised by the listener. It was thought that these invariant features are often hidden by the presence of other, less relevant, acoustic features or can be obscured by distorting the speech sound wave, but that the listener can detect them nonetheless and thereby recognise the phoneme sequence. It was said that automatic speech recognition could be achieved by detecting these invariants, always present although sometimes hidden, if only their nature was uncovered by further research and their characteristics specified. During the past decades considerable effort has been expended on finding these invariants. The development of sound spectrography or "Visible speech" (45) by the Bell Telephone Laboratories provided a most valuable means for the careful

analysis of speech sound waves and triggered off research at the Haskins Laboratories (3) (39) (40) (41) (43) and elsewhere. Although these studies have advanced our knowledge of the acoustic correlates of phonemes, etc. immeasurably, they did not produce evidence about the invariants: it is the thesis of the work described in this report that there are, in fact, no such invariants and that speech recognition is possible without their existence. Phoneticians are familiar with many examples in which, in fact, the same acoustic wave is recognised as one phoneme or as another, depending on circumstances. For instance, the words *man* and *men* are distinguished from each other by the vowel phoneme. It is known that the quality of this vowel varies considerably from speaker to speaker and that it is quite possible that for one speaker the vowel quality for *man* will be the same as another speaker's *men;* nevertheless when someone listens to these two speakers there will be no difficulty in distinguishing the singular from the plural. A more quantitative demonstration of the same effect is given by the results of an experiment in which listeners had to recognise synthetic speech (37). Different versions of a carrier sentence were synthesised in which the frequencies of the first two formants were varied; four syllables with different formant configurations were also synthesised. In the experiment each test item consisted of a carrier sentence followed by one or the other of the four syllables and the listeners were asked to identify the syllable as *bit, bet, bat* or *but*. The results showed that one and the same syllable was recognised quite differently depending on the range of formant frequencies used in the carrier sentence that preceded the particular presentation of the syllable. For instance, in the case of one of the syllables, when it was heard following a particular carrier sentence, it was recognised as *bet* 92% of the time and the judgments changed to 97% *bit* when the first formant frequency of the carrier sentence was raised. This experiment, then, provides further evidence to show that depending on circumstances one and the same acoustic event can be recognised as one linguistic unit or another and therefore speech recognition cannot be determined by these so-called invariant acoustic features. If there are no invariants, how does the listener recognise speech? This question may be answered best by re-examining the complete chain of events that takes place when a person speaks to and is understood by another; based on this it might then be possible to define how far a model of the human mechanism for speech recognition could be used as an automatic speech recogniser.

When a speaker wants to communicate with another person, he first organises whatever he wants to say in linguistic terms or in other words he formulates the information in the linguistic code. Language consists of a system of units that are combined into larger units according to rules peculiar to each language. For most purposes the phoneme is considered the smallest convenient unit, although the phoneme itself can be regarded as the combination of constituent elements, the distinctive features (34). The phonemes can be combined into larger units: the morpheme, the word, the sentence; each of the larger units can consist of one, or a combination of several, of the smaller units and represents a definite category of meaning. The number of the units varies from language to language. In English the number of phonemes is about 40; there are probably several tens of thousands of morphemes and words and the number of different sentences is very much larger still. During speech this linguistic code is transformed into a physiological one by the generation of a complex pattern of nerve impulses at various levels of the central nervous system; this pattern eventually produces a set of instructions that reach the muscles of the vocal organs via the appropriate motor nerves. The activity of these muscles produces movement of the vocal organs, the tongue, the lips, teeth, soft palate and vocal cords. The movement of the vocal organs generates the speech sound wave and brings about a transformation of the encoded speech information from a physiological code into an acoustic code. The sound wave reaches the listener's ears, stimulates his hearing mechanism and thereby generates a pattern

of nerve impulses in the acoustic nerve: this constitutes a re-conversion of the acoustic code into a physiological code. The nervous activity travels along to the central nervous system and eventually reaches the cortex of the brain and influences the nervous activity already there. The integration of the nervous signals arriving from the lower levels with the existing cortical activity somehow or other brings about the recognition of a sequence of linguistic units and eventually the understanding of the message. In the process of transforming the linguistic code into physiological and acoustic codes and back into a linguistic code the relationship between events on any two of the levels is by no means a one-to-one relationship. Also, whilst events on the linguistic level consist of a sequence of discrete units, the acoustic changes and also many of the physiological changes are continuous in nature. The transformation from the continuous acoustic signal to the sequence of discrete linguistic units requires quite a number of stages, but it is proposed here that they all belong to one of two main types of process: first, the assignment of a sound to one or other linguistic category based on the acoustic characteristics of the sound wave input, and secondly, the modification of this process of primary recognition by the statistical and structural constraints of the language. A considerable body of information is available about the acoustic cues on which the primary recognitions, that is the classification of all the incoming sound patterns into the framework of the 40 or so phonemic categories, are based. This large body of detailed factual data about cues for speech recognition are not only interesting in themselves but they also allow certain general conclusions to be made about the way in which primary recognitions are made. It has been established, for example, that often more than one cue may serve to identify a phoneme: the Haskins work has shown that plosive consonants are identified by the spectral position of the "plosive burst" and also by the nature of the formant "transition" of the adjoining vowel. Each of these cues on its own is sufficient to identify the plosive, but in normal speech both cues are discernible simultaneously: this type of redundancy is one of the reasons for the well-known fact that speech intelligibility is maintained even after severe distortion of the speech sound wave. Not all the cues for recognition are spectral, that is connected with the formant structure: intensity, fundamental frequency and duration may also play their part. Often the cues for recognition are not the absolute values of the acoustic signals along these dimensions, but their relative values. For example, turbulent energy at the end of an utterance will usually lead to the recognition of a fricative consonant in the final position; the duration of the turbulent energy can serve as a cue for classifying that sound as a voiced or unvoiced phoneme (5). It is not the absolute duration, however, that matters, but the duration relative to that of the preceding vowel sound: turbulence of a given duration nay be interpreted as a voiced or as an unvoiced fricative depending on whether the duration of the preceding vowel sound is long or short. Similarly, for fundamental frequency and intensity it is the relative rather than the absolute values that matter.

Although such extensive data are not available about the acoustic cues for phonemic classification, these cues do not provide anything in the nature of an invariant relationship between acoustic characteristics and phonemic class. The acoustic cues for a particular phoneme are widely scattered about some mean and the scatter is sufficiently large to produce considerable overlap into the acoustic areas of other phonemes. An additional complication is that there is no obvious way in which the acoustic sequence can be segmented to correspond to the successive phonemes. Even when divisions are introduced at the boundaries of acoustically dissimilar sections, it is often found that acoustic characteristics on both sides of the boundary have to be considered to identify a phoneme. If, for example, the spectrum of a syllable like /ni:/ is considered, a clear division into two segments

will be observed. In the initial segment, produced when the air was flowing through the nose, the sound intensity is low and there are only a few fairly broad formants. In the second and final segment the intensity is much higher, the formants much more clearly observable. On first sight one would want to identify the initial consonant with the first segment and the following vowel with the second segment. In fact both segments are needed to identify the nasal consonant: the initial segment only indicates that the consonant is a nasal one and the second segment indicates which of the three English nasal consonants is concerned.

What enables the human listener then to recognise speech with the high efficiency that he is usually capable of, if in fact there is no clear-cut relationship between the acoustic characteristics and the phonemes? The answer is that when the listener recognises speech he uses not only the acoustic information derived from the sound wave but also his knowledge of the subject matter and of the rules of the language used: a knowledge which he acquires over the years as the language is learnt. When recognising English speech, for example, the listener knows that he has to recognise phonemes in a system containing 40 units in all. He knows that these phonemes do not follow each other in any sequence but that certain sequences are much more likely than others. He also knows that the phonemes combine into morphemes and the morphemes into words. He knows all the possible morphemes and words in the language, the rules for combining them into sentences and also the ways in which expectations on the higher levels will affect those on the lower levels: once a number of words have been recognised, they will determine what set of words is most likely to follow and this information is fed back and will influence the sequential probabilities on the morphemic and phonemic levels. The strong effect of the constraints at the higher levels on the expectations at the lower ones can be demonstrated experimentally. Results of a test are available (22) in which the extent to which the subject could predict the phoneme sequence in a sentence was measured. The subject did not know in advance what the sentence was but had to guess it phoneme by phoneme; a record was kept of the number of times he had to guess before obtaining the correct answer for each of the phonemes in the sequence. The results show that about half of the first guesses were correct, demonstrating the strong effect of linguistic constraints. The results show further that the number of wrong guesses (the uncertainty) was greater for phonemes at or near the beginning of words and that the number of these wrong guesses at the beginning of words becomes less and less for words towards the end of the sentence, showing how the effect of constraints at higher levels is fed back to the lower ones.

Generally speaking then, when phoneme recognitions are made certain expectations are available which restrict the alternatives from which a choice is to be made when the acoustic information is received. Sometimes these expectations are so strong that the final choice can be made without acoustic data altogether. The realisation that such sequential constraints do in fact operate makes it possible to omit the vowels in the written form of certain languages: it was realised that once the consonants were written down the reader could guess the vowels by using his knowledge of the language. Although in English the vowels are written down, most readers would probably find no difficulty in understanding the following sentence in which only the consonants are written and all vowels are replaced by an x:

thx cxt sxt xn thx mxt

It may be worth while to give one more example of the effect of linguistic constraints on recognition. When conversation is carried on over a noisy telephone line it often happens that normal conversation is quite intelligible, but as soon as an unusual word or a proper name is mentioned it has to be spelt out to be understood. This shows that in quite normal situations the acoustic cues available are

not sufficient and linguistic information is essential for speech recognition. For the unusual words or proper names the constraints fed back from the sentence and word levels are not sufficient, the acoustic cues are not unambiguous enough and the recognition process breaks down. A similar example, of course, is the use in communication systema of the Able, Baker, Charlie spelling alphabet. When the lettera are spelt out as a, b, c, etc. the listener has to rely largely, though not entirely, on acoustic cues for recognition; when the words Able, Baker, etc. are used instead, then the English word system provides considerable linguistic constraints which make understanding less uncertain.

There is a considerable body of evidence, then, showing that there are no acoustic characteristics that have an invariant relationship with the phonemea and which although as yet unknown could be discovered by experiments: the human being does not rely solely on acoustic cues for speech recognition but utilises strong constraints arising out of the linguistic organisation of the transmitted information. If even a human being cannot recognise speech successfully by using acoustic criteria alone, then an automatic speech recogniser is not likely to be able to do so either, and it seems worth while to investigate the use of linguistic statistics for automatic recognition. This has been done by constructing an automatic apeech recogniaer which utilises both acoustic and linguistic information.

Before embarking on the detailed design of the recognition system, a general question has to be settled: the kind of linguistic unit in terms of which the recognitions are to be made. It has already been pointed out that an easential feature of the human recognition process is that the acoustic input is classified into a restricted number of basic units and that these are then combined into larger units. The automatic recognition process has to perform an analogous function but in theory the unit of recognition need not necessarily be the phoneme, it could be one of the larger units and the choice must be made on the grounds of convenience. The larger the unit the greater the number of individual units in the system: there are only about 40 phonemes, but many thousands of words and an even greater number of sentences. The requirements for storing linguistic information in a machine dealing with phonemes are therefore considerably more modest than in a word recogniser. This is even more so when the question of storing the statistical information about the probability of sequences of these linguistic units is considered: the possible number of sequences of four phonemes is around 2½ million whilst the theoretical maximum of just two-word sequences is about a hundred million. At first sight this implies that the recogniser would have to store a much larger number of items if the word is chosen as the unit of recognition than if the phoneme is chosen. This need not necessarily be so because it may well be that primary recognition of words is more successful than that of phonemes and therefore less linguistic information may be needed to achieve the same overall success than with a phonemic system. Furthermore, a large proportion of phoneme and of word sequences never occur at all and it is not known how many significantly different values of phoneme and word sequence probabilities would have to be stored for equally successful recognition. In the light of future evidence, it may be found therefore that a word-based recognition system is more economical in storage requirements than an equally successful phoneme recogniser, but for the purposes of the present work the phoneme seemed to be by far the most economical unit. In this way the number of basic recognition units could be kept quite small, to just a selection of phonemes, and the recogniser could still deal with speech material consisting of several hundred words and at the same time requiring memory capacity for only a few hundred values of phoneme digram frequency. Had the word been used as the basic recognition unit, then if the system was to deal with the aame number of words as before, the memory of digram frequencies would have had to be made much larger (although, of course, the system might also work more efficiently).

CHAPTER IV

THE DESIGN AND CONSTRUCTION OF THE AUTOMATIC PHONEME RECOGNISER

In the light of the foregoing discussion it was decided to construct an automatic phoneme recogniser which used both acoustic and linguistic information in its operation

The general scheme for such a system is shown in Fig. 5. The speech sound wave is

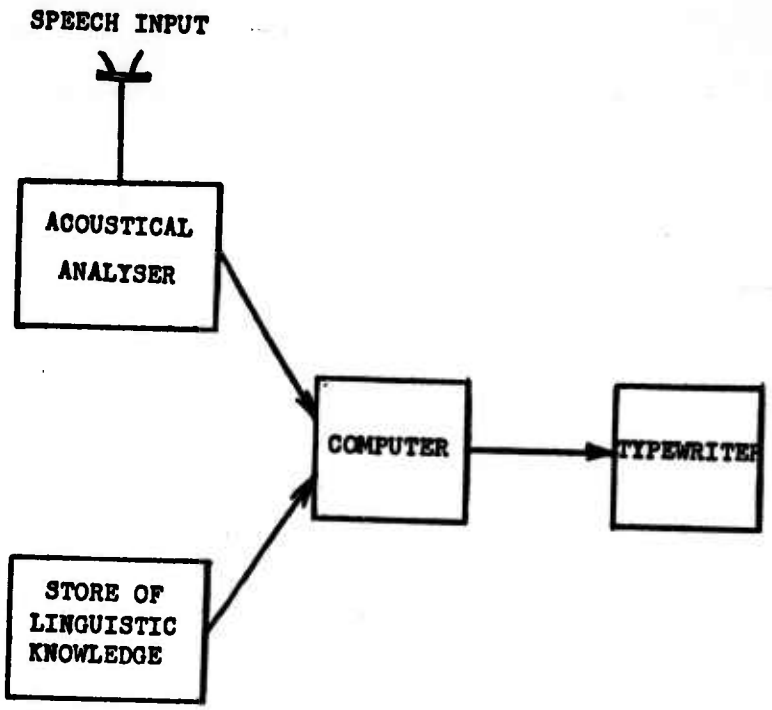


Fig. 5. Block diagram illustrating principle of operation of automatic phoneme recogniser.

applied to the acoustic analyser where it is examined in various ways and a preliminary phonemic classification, based on relevant acoustic characteristics, is then indicated at the output. In another part of the recogniser, the "store of linguistic knowledge", information is available about the probability of occurrence of the various linguistic units as a function of context. In the case of the present recogniser, this linguistic information is in the form of digram frequencies, that is the probability of occurrence of the various linguistic units in the repertory of the machine as a function of the identity of the immediately preceding phoneme. The final recognition is made by the "computer" which combines the information derived from the acoustic analyser and from the store of linguistic knowledge. The computer makes its final decision then, by considering the acoustic cues derived from the speech sound wave and the sequential probabilities determined by language statistics. The decisions of the computer are

indicated on an electric typewriter. The typewriter was used purely as a convenient way of recording the output of the computer rather than as part of an effort to construct a "speech typewriter", a device which types out the speech spoken into it. The principal aim of building the recogniser was to see what improvements, if any, could be achieved when linguistic statistics were used to modify the results of acoustic recognition, rather than to pursue the design of the acoustic detector to a great degree of refinement or to achieve a practical automatic recogniser as such.

The machine to be described here was designed to recognise English words, spoken in isolation and using Southern English pronunciation. The method of selecting the list of test words will be described later. For the sake of economy of construction the phoneme repertory of the recogniser was restricted to 12 phonemes: 4 vowels, 7 consonants and the space between words which was also treated as a separate phoneme as it had a definite structural function. The vowel phonemes were chosen to be as representative of English vowels as possible. /i:/ as in *heat*, /a:/ as in *heart*, /u:/ as in *hoot* were selected as being the three English vowels whose articulation was nearest to the three corners of the "vowel triangle" and / ɘ:/ as in *hurt* was selected because its articulation corresponds to a position near the centre of the vowel triangle. It was well-known that the recognition of the consonants was more difficult and, while trying to keep the selection as representative as possible, the more difficult ones were not necessarily included. In particular, while representatives of every "manner of articulation" used in English were included, on the whole those phonemes which are articulated with relatively high intensity were used; for this reason the consonants were as far as possible of the unvoiced variety. The seven consonants were /t/ and /k/, /s/, and /ʃ/, /m/ and /n/, and /l/. Later the repertory of the recogniser was extended to include two additional consonants, /z/ and /f/.

The test words were recorded on magnetic tape and in all experiments the playback from the tape was used as the input to the recogniser instead of live speech. Care was taken to use a recording system with as good a signal-to-noise ratio as possible in order to accommodate the very wide dynamic range of speech sound waves. An Ampex 600 was used for recording and also for play-back. The recorder was modified to run at 15 in./sec. and full-track heads were used. In this condition it gave a signal-to-noise ratio of 60 db. as measured on the screen of a cathode ray tube. The overall frequency response, comparing the electrical input to the recorder with the electrical output from the play-back, was flat within ±1 db. over the range 60 c.p.s. to 15,000 c.p.s. A Standard Telephones and Cables type 4021F moving coil microphone was used which has a comparably good frequency response. The playback amplifier was followed by pre-emphasis, a certain amount of peak clipping and power amplification. Pre-emphasis was used to equalise the average speech spectrum which otherwise would be falling towards the higher frequencies. The pre-emphasis amounts to about 4 to 5 db. per octave over the frequency range 500 to 4000 c.p.s. The circuit is shown on the left of Fig. 6. The signal from the tape recorder is connected to valve V_1 which is used as an anode follower with a 0.005μF capacity across the input resistance. The output of the pre-emphasis circuit was peak-clipped to provide a crude form of volume compression; spectrographic analysis showed that the distortion produced by the peak clipping did not materially modify the speech spectrum. The clipping amounted to 20 db. relative to the peaks of the speech wave as observed on the screen of a cathode ray tube. The circuit can be seen in Fig. 6 and is centred on the diodes D_1 and D_2. When no signal is applied the live end of the 2.7 K ohm resistance receives a bias of about + 5 volts from the positive H.T. line. Any signal arriving from V_1 is then clipped symmetrically at the ± 2.5 volt level. The output of the clipper is applied to the power amplifier, consisting of valves V_3 V_4 V_5 V_6. The output available is 8 watts into 300 ohms with under 1% distortion. The output

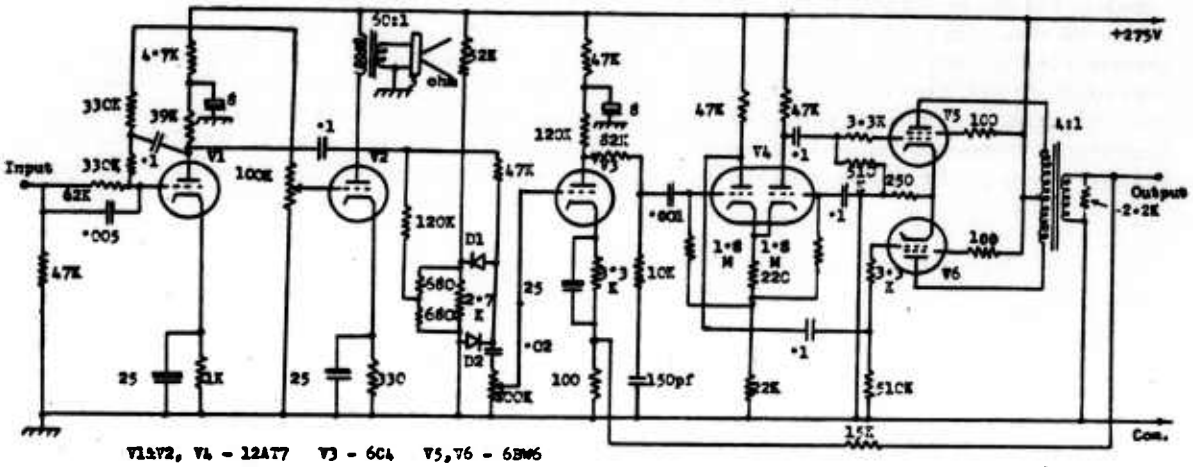


Fig. 6. Circuit diagram of filter-bank amplifier.

On this and all other circuit diagrams the values of resistance are shown in ohms and of capacitance in microfarads, unless otherwise stated.

of this power amplifier provides the input for all circuits concerned with scoustic recognition.

The design of the acoustic recogniser was based on the detection of well-established acoustic cues only because, as pointed out earlier, the purpose of the acoustic detector was to carry out a simple recognition process which could then be modified by linguistic information; the effect of using this linguistic information could probably be evaluated even if the detection of acoustic cues is not the most perfect in the light of what is known on the subject.

All acoustic recognition processes were based, partly at least, on spectral cues and the necessary spectral analysis was obtained by applying the speech sound wave to a bank of filters. The filters covered the range from 160 c.p.s. to 8000 c.p.s. in 18 adjacent bands, with each filter about a third of an octave wide and with their mid-band frequencies spaced about a third of an octave apart throughout the range. Each filter consisted of an inductively coupled, double section, series tuned circuit as shown in Fig. 7. At resonance the output voltage was about 3.8

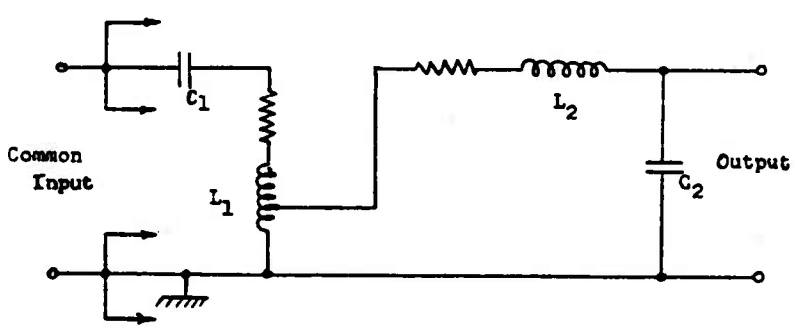


Fig. 7. Circuit diagram of analysing filters.

times greater than the input. The frequency response curves for three adjacent sections shown in Fig. 8 are typical of all the filters. It will be seen that the mid-frequencies of adjacent filters are a third of an octave apart, 6 db. attenuation is obtained at the cross-over points, about 15 db. attenuation at the mid-frequency of the adjacent filter and about 35 db. an octave from the mid-frequency. The mid-band frequencies and corresponding serial numbers of all 18 filters are also shown on this graph. Such a logarithmic distribution of bandwidths and of mid-frequency spacing was

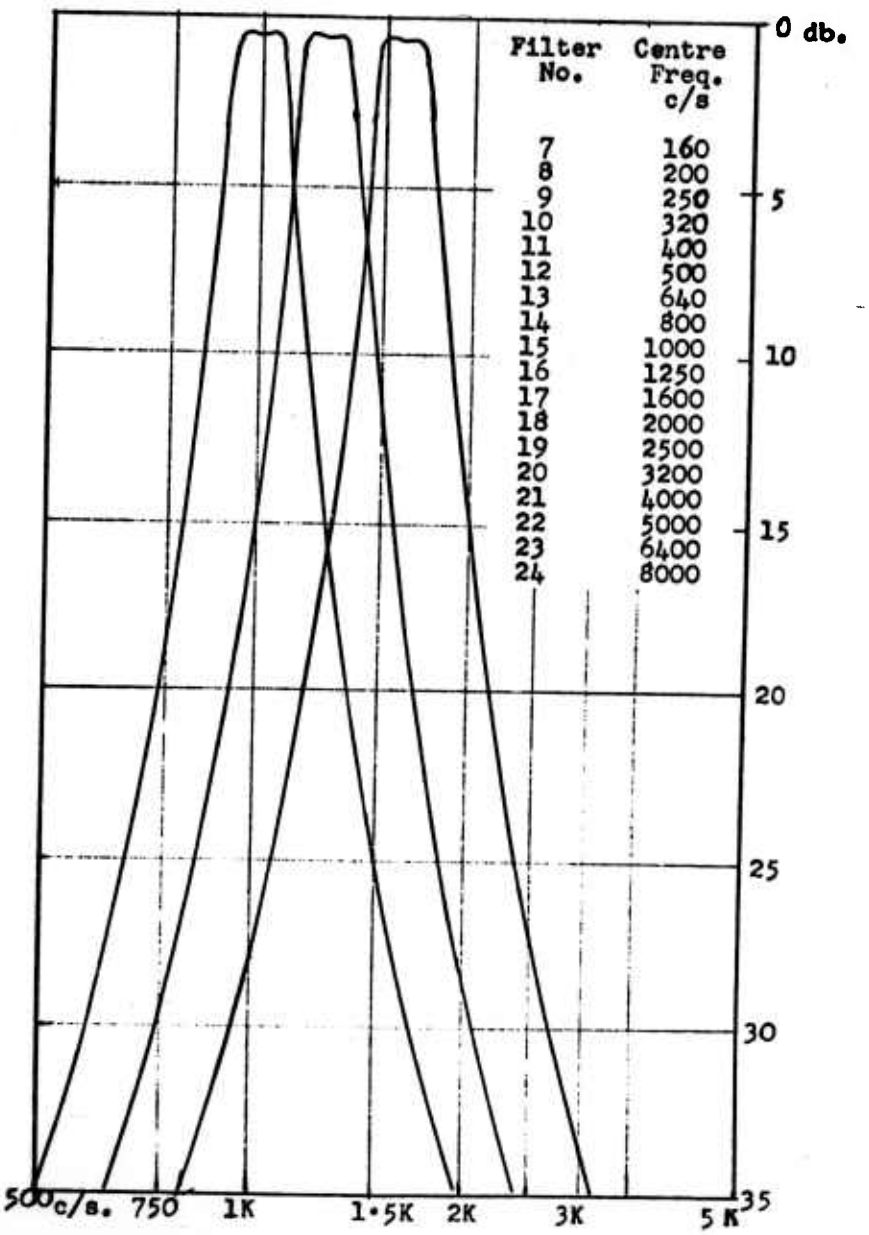


Fig. 8. Frequency response curve of three typical filters (numbers 15, 16 and 17) and serial numbers with centre frequencies of all 18 filters.

chosen because this psrticulsr filter bsnk was avsilable at the time the work was started rsther thsn becsuse it was thought to be the most suitable for speech wave analysis. It would hsve been prefersble to have filter spscings and bandwidths that were constant at about 100 to 150 c.p.s. up to about 1 kc.p.s.or 2 kc.p.s. and a logsrithmic increase for the higher frequencies. Such s distribution would correspond to equal steps on the "Koenig" scsle (36), s scsle specially designed for the analysis of speech waves. In the filter bsnk used, st the lowest frequencies the bsndwidth is so nsrrow thst the time constant is too long to follow fast varistions of energy that msy be significant cues for speech perception, while in the 1 kc.p.s. to 2 kc.p.s. region the bandwidth is too wide to distinguish significant varistions of formant frequency.

Esch filter is followed by s rectifier and smoothing filter to obtsin s measure of the energy level in esch frequency band. The rectifier circuit is shown in Fig. 9. The values of condensers C_1 C_2 C_3 were different for the different filter

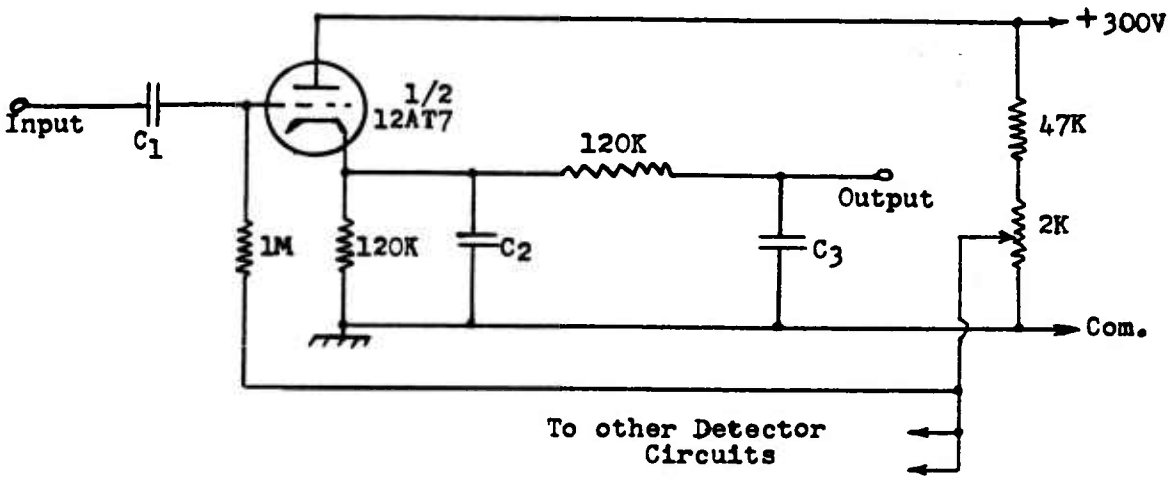


Fig. 9. Circuit diagram of filter rectifier and smoothing filter.

channels. Their vslues were kept ss small ss possible consistent with ressonsble smoothing (not more thsn 10% ripple) and a low psss cut-off frequency (3 db. point) not less thsn 100 c.p.s. or the bandwidth of the preceding filter channel, whichever was lower. The 2000 ohm potentiometer shown on the right-hand side of the disgram provides a permanent bias of + 8 volts st the output of all filter rectifiers; this was found convenient for the multiplier circuits to be described lster. The input-output curve, showing the relationship between the peak A.C. input and the D.C. output voltsge wss linesr to 1% over the rsnge used (8-100 volts D.C. output).

Turning now to the sctual sutomatic recognition processes, the many different acoustic cues thst sre svsilable for sutomstic recognition have been discussed previously in publications (24) and in this report on the whole only those cues will be mentioned that have been utilised in the recognition processes described here.

THE AUTOMATIC RECOGNITION OF THE VOWELS

All vowels and some of the consonants were recognised by detecting characteristic spectral envelopes. It has long been acknowledged that the most characteristic feature of vowel sounds is the frequency of the formants and the frequencies of the maxima of the spectral envelope are more often than not a close approximation to these formant frequencies. The relevant spectral peaks were determined in preliminary experiments. This was necessary because no published information is available on the average formant frequencies of British English (although data are available for American English (44)). It would have been desirable to re-examine the structure of spectral peaks for the speech material to be used in the automatic recognition experiment even if formant data in the form of average values for English had been available. This is partly because average values can differ greatly from individual results for particular phonemes and partly because the selectivity of the filters used for analysis will influence the extent to which individual formants can be isolated. The recorded speech material was therefore applied to the bank of filters and the rectified and smoothed output voltage of each filter was recorded using a multi-channel pen recorder. Typical records are shown in Figs. 10(a) and 10(b). The output of only 6 different filters and for only a

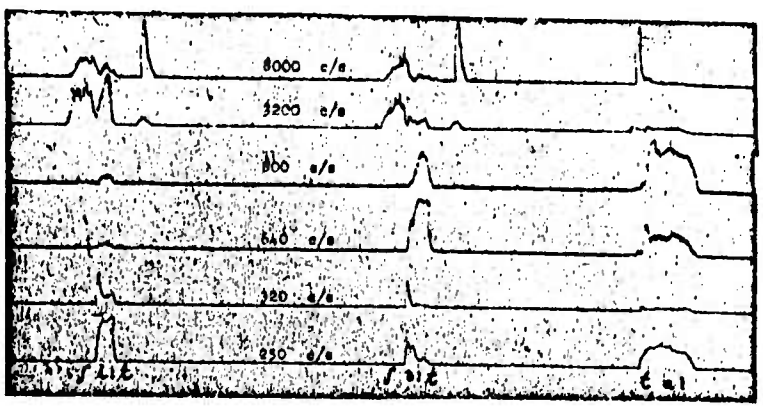


Fig. 10 (a).

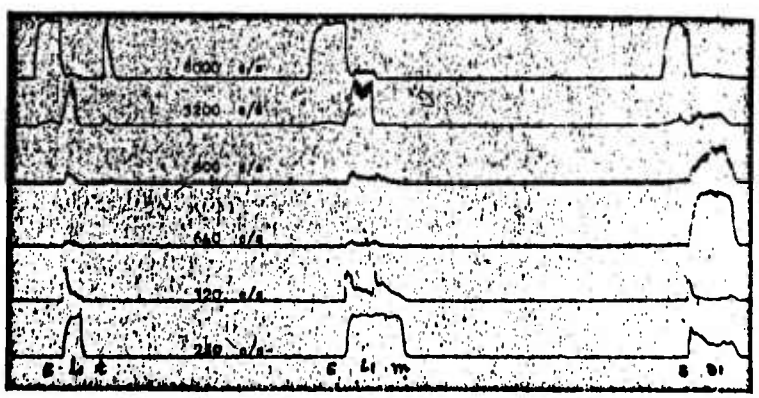


Fig. 10 (b).

Fig. 10. Pen recording of rectified filter outputs for a few test words.

few test words are shown in these examples; in the preliminary experiments all 18 filter outputs for all the test words were displayed side by side in this fashion. The change in the spectral distribution of energy as the vowel changes from /i:/ to /ə:/ to /a:/ can be seen clearly in Fig. 10(a), and for /i:/ and /ə:/ in Fig. 10(b). By examining records of this kind it was found that the four vowels represented in the speech material of the input could be adequately differentiated by specifying two filter channels in which energy peaks occurred simultaneously. These peaks were in the filters centred at 400 c.p.s. and 800 c.p.s. for /u:/, at 250 c.p.s. and 3,200 c.p.s. for /i:/, at 640 c.p.s. and 1,600 c.p.s. for /ə:/, and at 640 c.p.s. and 1,250 c.p.s. for /a:/. The values of frequency indicate that it is mostly the first and second formants that determine these peaks, although occasionally, as in the case of /i:/, it is the third formant or the merging of second and third formants in one filter band that are responsible for the spectral peak. Automatic recognition of the four vowels was obtained by comparing the spectrum of the speech input with the four spectral patterns specified in terms of the above pairs of spectral peaks and indicating with which of these four the input corresponds best. The method used to achieve such a comparison was to multiply the rectified output voltages of the appropriate pairs of filters; in this way as many multiplication products become available as there are phonemes to be recognised - four as described so far. The products are then compared in magnitude and the largest selected. A schematic diagram of such a system is shown in Fig. 11. This system of spectrum-matching has

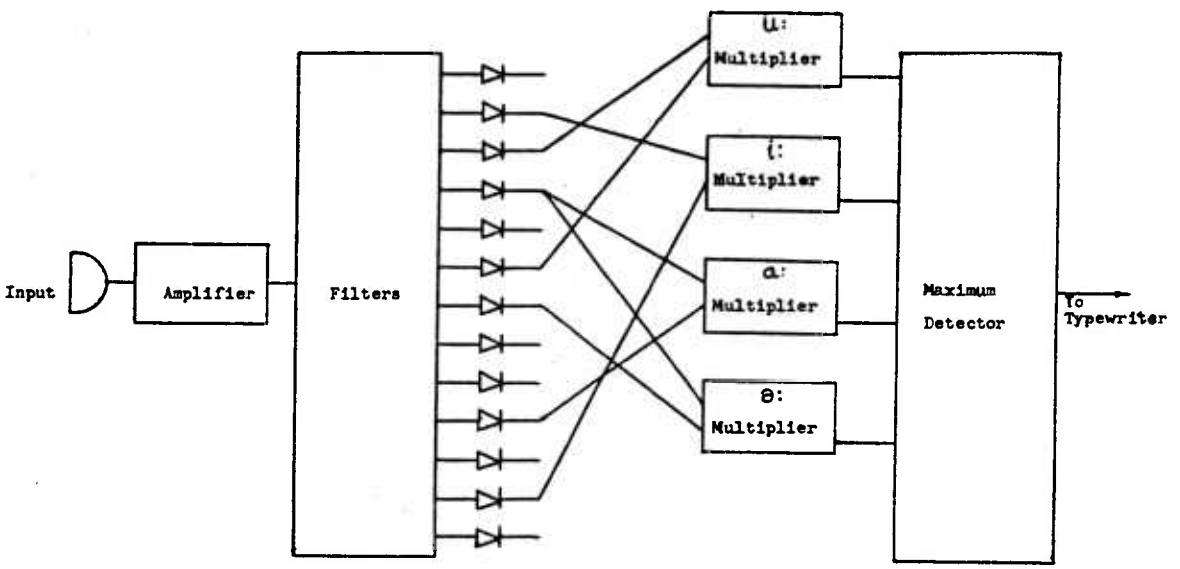


Fig. 11. Schematic diagram of spectral pattern matching device.

several advantages. The incoming spectral patterns are necessarily assigned to one or other of the categories of spectral configuration which makes up the predetermined system within which the automatic recogniser operates. Another advantage is that no threshold judgment is involved: the recogniser compares the spectrum of the input with each of the reference patterns and selects the best match, rather than basing its judgment on whether the combined energy from a given pair of filters exceeds a certain level or not. This method also provides quite a reasonable solution of the "segmenting" or "gating" problem: whenever the spectrum of the input changes sufficiently for the best match to shift from

one reference pattern to another a "phoneme boundary" is indicated. Yet another advantage is that not only is the best match indicated at the output, but information is also given about how closely the input resembles the other reference patterns. This information is available through the output voltages from the other multipliers and is going to be of great advantage later when the results of acoustic recognition and linguistic information are combined to give a final recognition. Fig. 11 shows that the principal circuit elements of the pattern matching process are the multipliers and the maximum detector and these will now be described in more detail.

The multiplier circuit tried out in the first instance used a Mazda 6F33 type of valve. This is a pentode with a suppressor grid which has the same order of sensitivity as the control grid. The two voltages were applied to the control grid and suppressor grid respectively and the output was obtained from the anode. This type of circuit however was soon abandoned, partly because of its instability and partly because of its non-linearity. The next circuit that was tried, the one which was eventually put into use in the recogniser, carried out the multiplication process by generating a square wave in which the mark-space ratio was controlled by one of the voltages to be multiplied and the amplitude of the square wave was determined by the other voltage to be multiplied. The area underneath such a square wave is proportional to the product of the height and length of the square wave and hence to the product of the two input voltages. An output proportional to this area is obtained by integrating the square wave. The square wave necessary for the multiplication process is derived from a triangular wave, generated by the circuit of Fig. 12.

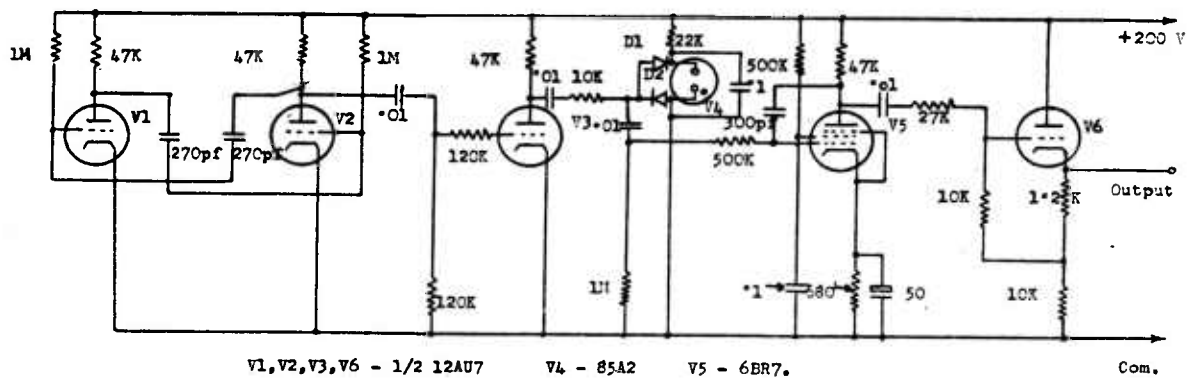
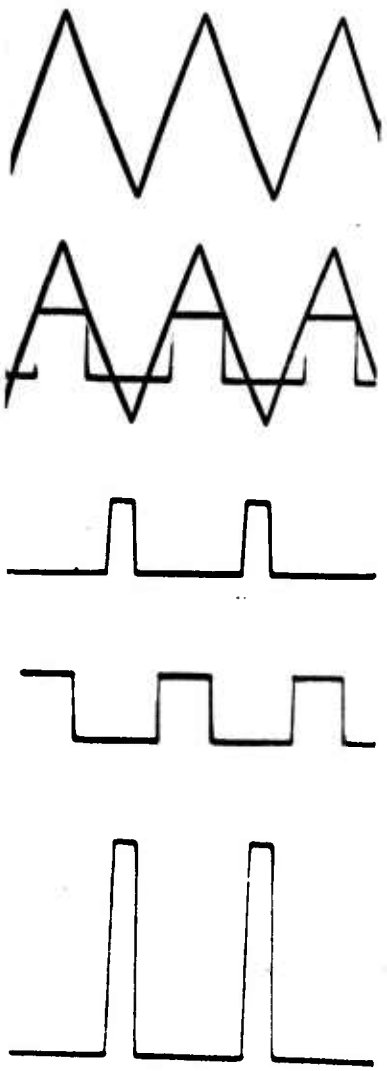


Fig. 12. Circuit diagram of triangular wave generator.

The output of a free-running multivibrator circuit, incorporating valves V_1 and V_2, is amplified and squared by valves V_3 V_4 D_1 D_2. The square wave output is applied to the Miller integrator of V_5 and thus provides a triangular wave which is connected to the final output through the cathode follower V_6. This triangular voltage, the wave shape of which is shown in Fig. 13(a), is used with all the multiplier circuits. The principle of obtaining the variable square wave, necessary for the multiplication process, from the triangular voltage is shown in Fig. 14 A thin slice of the triangular voltage is cut out and the mark-space ratio of the resulting square wave is determined by the slicing level which in turn is set by one of the voltages to be multiplied. This square wave is amplified and is then amplitude-limited at a level determined by the other voltage to be multiplied.



(a) Waveshape of triangular voltage generator.

(b) Composite oscillogram showing triangular voltage and a square slice from it.

(c) and (d) Examples of mark-space ratios obtained when slicing at different levels.

(e) A square wave with the same mark-space ratio as in (c) but limited to a different amplitude.

Fig. 13. Oscillograms demonstrating the operation of the multiplier.

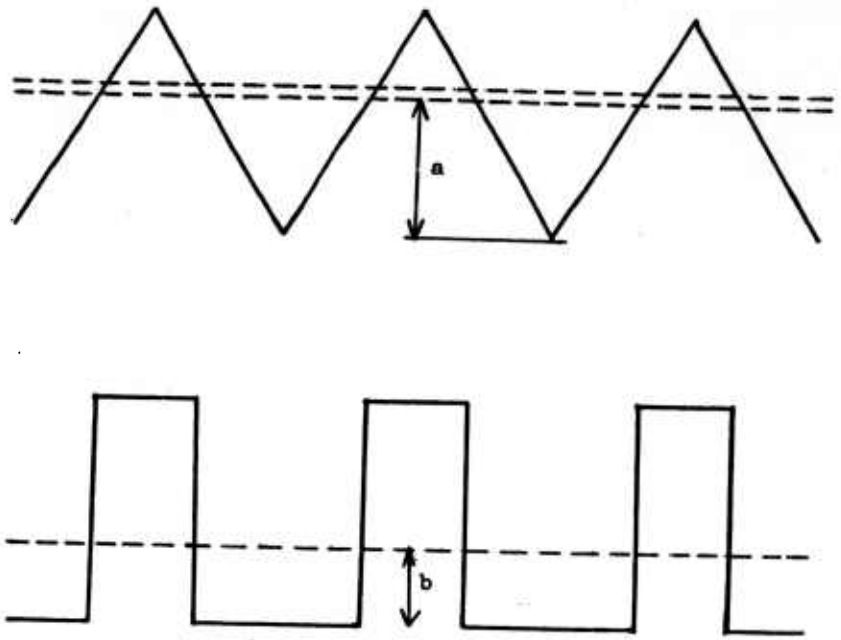


Fig. 14. Principle of operation of multiplier.

The corresponding circuit diagram is shown in Fig. 15. One of the voltages to be multiplied is applied to the grid of the buffer cathode follower valve V_1. The output from V_1 is mixed at the grid of V_2 with the triangular voltage. The combined

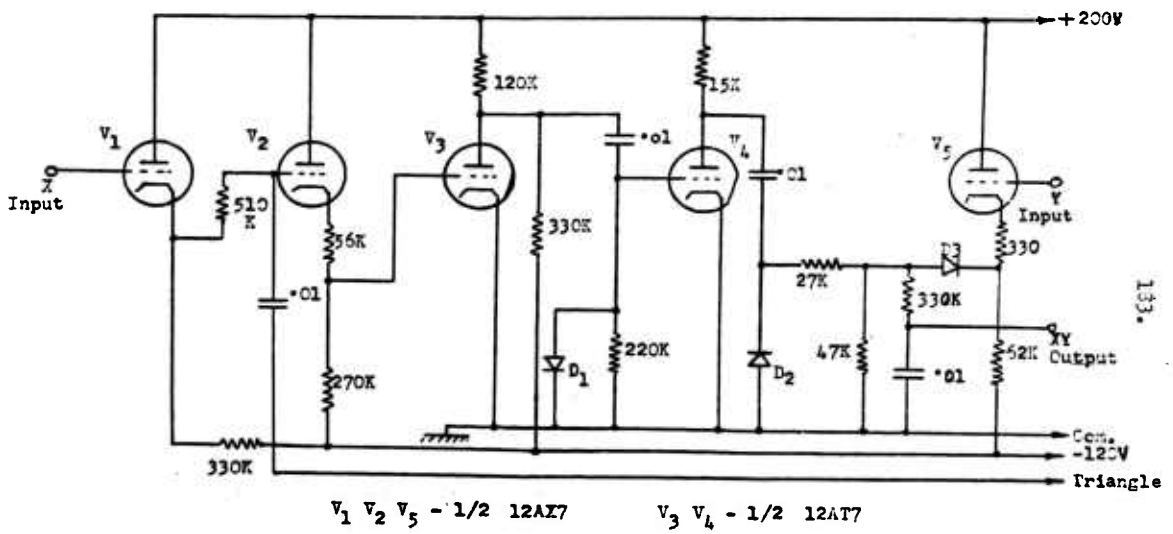


Fig. 15. Circuit diagram of a typical multiplier section.

voltage is in the form of a triangular voltage whose absolute level is determined by the input from V_1. Valves V_3 and V_4 perform a squaring and amplifying function, thereby cutting a square slice from the triangular voltage at a height determined by the voltage derived from V_1. The amplified square wave is then amplitude-limited at a level determined by the other voltage to be multiplied which is connected to the grid of V_5: the amplitude of the square wave from V_4 cannot rise above that of the cathode of V_5 because of the clamping diode D_3. The diode D_2 D.C. restores the output of V_4 to the zero voltage level. The square wave is then integrated by the 330 K ohm resistance and the 0.01 uF condenser and the integrated voltage from the condenser provides the output voltage proportional to the product of the two voltages applied to the grids of V_1 and V_5 respectively. Typical wave shapes illustrating the multiplying process are shown in Fig. 13. The frequency of the triangular voltage had to be high compared with the rate of change of the voltages to be multiplied. These input voltages are derived from the speech spectrum and no significant change can reasonably be expected in less than 1/20th to 1/30th of a second. For this reason the frequency of the triangular voltage was set around 3,300 c.p.s. so that at least 100 cycles of this voltage take place during the shortest significant period of the voltages to be multiplied. The time constant of the integrating circuit is 3.3 msec., that is about ten times the period of the triangular voltage, and the time constant of the multiplier as a whole, as measured on the screen of a cathode ray tube, is about 6msec. This time constant, which is long relative to the triangular wave period and short relative to the speed of variation of the multiplying voltages, ensures that the output has no noticeable 3,300 c.p.s. ripple but still follows the changes of the input voltages. The linearity of the multiplying action was checked by connecting both inputs to the same voltage and obtaining a curve relating this common input with the square root of the corresponding output. The resulting curve was a straight line over the whole of the operating range of 20 volts. The stability of the circuit was good. The H.T. was supplied from a stabilised power pack which is described at the end of the section dealing with the circuitry of the automatic recogniser. This stabilised power pack not only kept the H.T. voltage constant to better than 1% but also its output impedance of under 0.1 ohm over the whole of the relevant frequency range ensured freedom from interference from other circuits. The heaters were not stabilised but did not affect the output: any effect due to heater variation would influence all the multiplier circuits equally and would therefore not affect the final result. The second part of the spectral pattern matching arrangement is the maximum detector and both a simplified and a more detailed circuit diagram are shown in Fig. 16. The voltages to be compared, derived from the multipliers and all positive going, are applied each to the grid of a different triode as shown in Fig. 16(a). The triodes have a relay in their anode circuits and they have a common cathode resistance. This resistance is given such a value that the current flowing through it can operate one only of the relays in the anode circuits. Whichever input is the most positive will divert the largest share of the current from the common cathode resistance to its own triode. At the same time, the cathode of this valve will, due to cathode follower action, go more positive and, as all cathodes are connected together, will cut off all the other valves. In this way the valve with the most positive voltage on its grid will capture most of the current available from the cathode resistance and the relay in its anode circuit will operate while all other relays release. The minimum amount by which a voltage to be compared has to be more positive than all the others is the difference in grid voltage corresponding to the operate and release currents of the relays. This is between one and two volts for the circuit which was used and was not considered sufficiently small compared with the 20 volt range of output voltage from the multipliers. Each input to the maximum detector was therefore amplified about seven times before it was applied to the appropriate grid of the comparison circuit. The detailed circuit is shown in Fig. 16(b). Valves V_1 V_2 and V_3 form one cell of the maximum detector circuit and are repeated 16 times to allow 16 different voltages to be compared. V_1 is a cathode

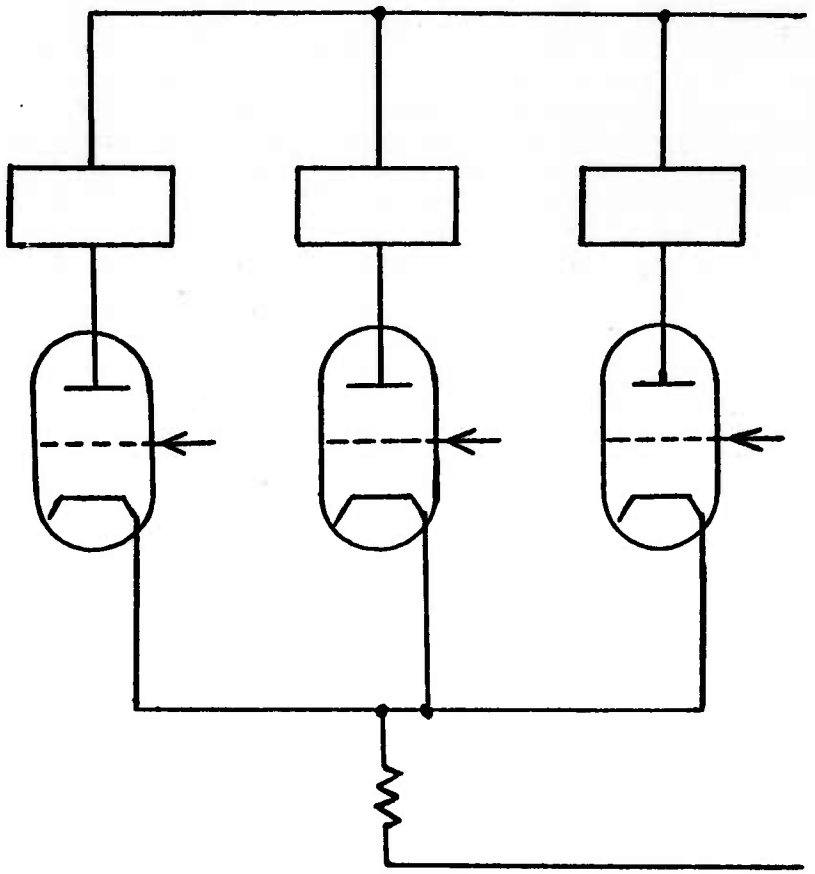


Fig. 16 (a). Maximum detector circuit. Simplified diagram.

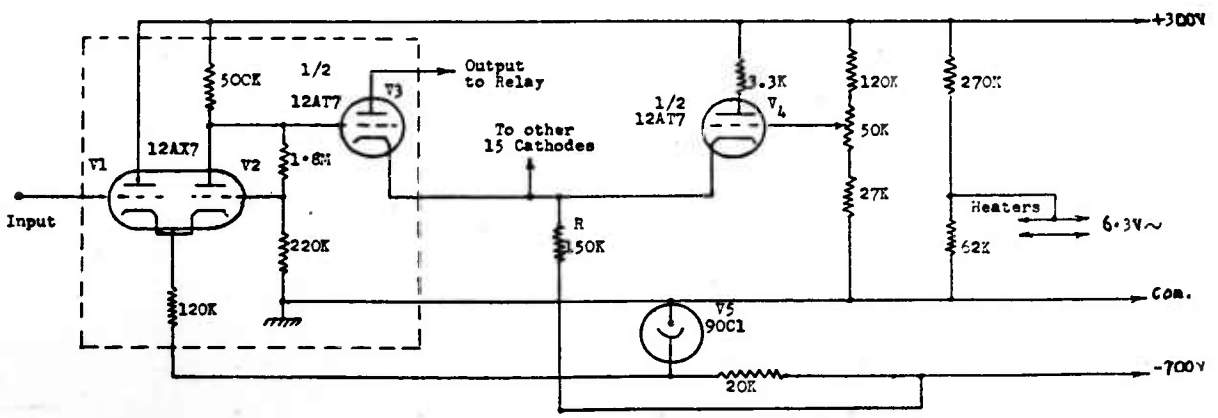


Fig. 16 (b) Maximum detector circuit. Detailed circuit diagram of one of 16 identical sections.

follower input stage which is cathode coupled to amplifier valve V_2. The gain of this stage is stabilised by strong feed-back from anode to grid. Further stabilisation is obtained by the cathode coupling between V_1 and V_2. The overall gain of V_1 and V_2 is about seven. V_2 is directly coupled to the grid of the comparison valve V_3. The cathode of V_3 is common with the cathodes of the other 15 V_3 valves and is connected to the negative end of the H.T. supply through the common cathode resistance R. The anode of each V_3 is connected to the positive end of the H.T. supply through the operating coil of a separate relay not shown in this diagram. The discrimination sensitivity of the circuit has been increased to about 0.25 volt referred to the grid of V_1 because of the gain of seven provided by V_1 and V_2. The circuit is extremely stable. The H.T. supply is stabilised and variations of heater voltage are neutralised by feed-back and also by the fact that they are likely to affect all 16 circuits to the same extent and will therefore not influence the maximum selection. Whenever no speech input is applied to the recogniser all the input voltages to the maximum detector become zero: no maximum is evident and therefore random operation might result. This is avoided by adding a 17th valve to the chain of V_3 valves. The grid of this valve, V_4, is connected to a fixed D.C. voltage of 54 volts whilst the voltages at the grids of the 16 V_3 valves vary from 48 volts when no input is applied to the maximum detector to about 190 volts for maximum input. This means that when all input voltages drop below about the last 4% of their total range, then the current is captured by V_4 and random operation is avoided. The result of the maximum selection is recorded by arranging that the closing of the contacts of a relay will operate one of the keys of an electric typewriter. The mechanical and electrical arrangement for the operation of the typewriter will be described after all the actual recognition circuits have been discussed.

THE AUTOMATIC RECOGNITION OF THE CONSONANTS /m, n, l, s, ʃ/.

All four vowels in the repertory of the automatic recogniser were detected by using the pattern matching process - with quite a high rate of success as will be seen later. When examining the spectrographic records, by using the multi-channel pen recorder as before, it was found that the same method could be used for the recognition of a number of continuant consonants, in particular /m/, /n/, /l/, /s/ and /ʃ/. As would be expected, reference patterns with peaks in the low frequency region were needed to recognise the nasal and lateral consonants and in the high frequency region for the fricatives. The filters centred on 200 c.p.s. and 320 c.p.s. were used for /m/, the 250 c.p.s. and 400 c.p.s. for /n/ and the 400 c.p.s. and 500 c.p.s. filters for /l/. These pattern allocations were rather tentative and, as will be seen later, these three sounds were never well distinguished from each other; it was easier to distinguish initial /m/ and /n/ sounds from final /m/ and /n/ sounds than it was to distinguish /m/ from /n/. Also quite different acoustic structure was found to apply to initial and final /l/. The initial /l/ was almost indistinguishable from /i:/ whilst the final /l/ produced a single broad peak in the 400 to 500 c.p.s. region. The /ʃ/ produced a broad peak in the 2 to 4 Kc.p.s. region whilst the /s/ energy was higher than this, around 6 to 8 Kc.p.s. The intensity of the fricative energy was high for both fricatives so that there was no difficulty in detecting the presence of the sound, although the usual double input to the multipliers was very necessary to establish that the energy was really concentrated in the appropriate spectral regions. If the presence of energy in only one channel had been taken as the criterion for recognition, then it would have been easy to confuse /ʃ/ with some of the vowels like /i:/ or /ə:/ which have high frequency formants overlapping into the range of /ʃ/. Similar difficulties would have arisen in distinguishing /s/ from /ʃ/. Some of these points can be observed in the pen recording shown in Fig. 10. For instance the shift of fricative energy for /s/ and /ʃ/ can be seen when comparing /si:t/ and /ʃi:t/. The overlapping of spectral cues for /ʃ/ and /i:/ can be seen on the tracing for /ʃi:t/.

In this way the first version of the automatic recogniser which was extensively tested had nine multiplier circuits for carrying out spectral pattern matching. Four of these were used for the vowels /a:/ /i:/ /u:/ /ə:/ and five for the consonants /m/ /n/ /l/ /s/ /ʃ/. The

table below gives the serial numbers and centre frequencies of the filters connected to each of these multipliers.

Phoneme	Serial number of filter outputs utilised	Centre frequency of filter outputs utilised c.p.s.
a:	13 and 16	640 and 1,250
u:	11 and 14	400 and 800
i:	9 and 20	250 and 3,200
ə:	13 and 17	640 and 1,600
m	8 and 10	200 and 320
n	9 and 11	250 and 400
l	11 and 12	400 and 500
s	23 and 24	6,400 and 8,000
ʃ	19 and 21	2,500 and 4,000

THE AUTOMATIC RECOGNITION OF THE CONSONANTS /t, k/.

Some of the phonemes selected for the repertory of the recogniser could not be detected on spectral cues alone. The acoustic characteristics of the phoneme /t/ for instance were the silence during the stop segment of the articulation and a short burst of hiss energy generated during the release of the stop. The silence is difficult to detect unless the /t/ is produced between two other sounds and is obviously quite unsuitable as a cue for recognising an initial /t/. The short fricative burst has a spectrum that extents over roughly the same range as the spectrum of /s/ and can be seen clearly in the pen recordings of Fig. 10. Although the burst has the same spectrum as /s/, its duration is always recognisably less than for /s/ and the detection of this difference in duration was used for recognising /t/. A similar procedure was used for recognising /k/ except that the spectrum of the burst was at a lower frequency, near that of /ʃ/ or even lower.

The circuits used for recognising /t/ and /k/ were therefore practically identical - they both used a form of duration measurement - and they differed only in obtaining their inputs from a different filter. The basic principle of the plosive detectors is that an electronic clock, set to measure off a fixed time interval, is started as soon as a voltage appears at the input of the detector circuit. If the input voltage disappears again during the span of time measured off by the clock then a plosive consonant is indicated; if the input is still present at the end of the measured period then the plosive detector becomes inoperative again without giving an output and the recogniser chooses its output from alternatives suggested by one of the multiplying circuits. A simplified diagram of the /t/ detector circuit is shown in Fig. 17. The output of the 8,000 c.p.s. filter is applied to the grid of the triode V_1 and will operate relay A as soon as the input from the filter exceeds a certain threshold value. This threshold is determined by the bias on the cathode of V_1, obtained from the 15 K ohm and 2 K ohm resistors, and by the setting of the input potentiometer. The 1M ohm resistor and 0.0015 uF capacitor prevent spurious operation of the relay by voltage spikes of short duration that sometimes appear at the output of the filter but are not connected with the phoneme /t/. The operate and release times of relay A and of all the other relays used in this circuit are less than 2 msec., fairly short compared with the duration of other significant events in the operation of the circuits. The triodes V_2 and V_3 form a one-shot multivibrator circuit in which V_2 is normally cut off and V_3 is normally on. This means that relay B is normally operated; therefore its contacts are normally in position 3 and relay C is not energised. An output of large

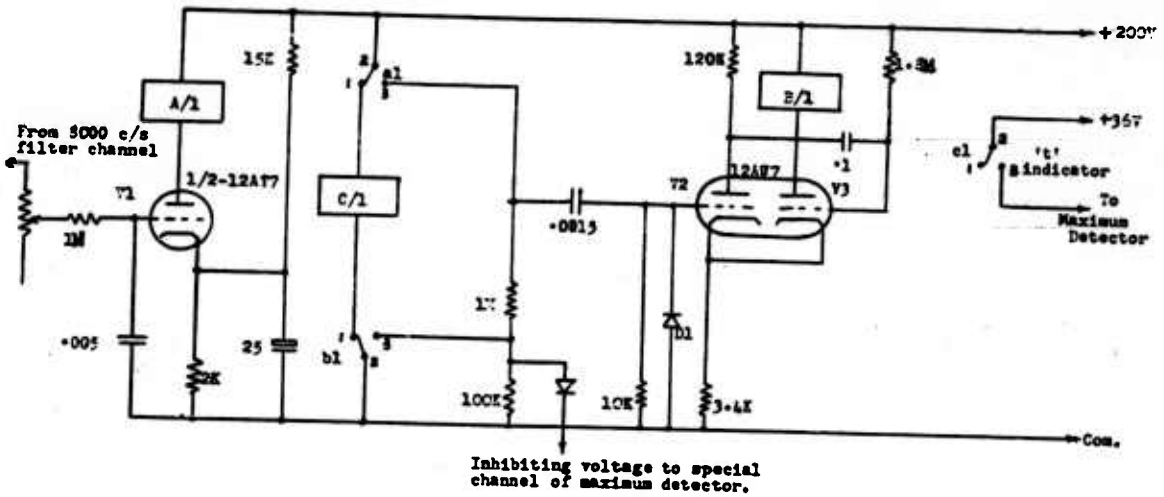


Fig. 17. Simplified circuit diagram of the /t/ detector.

enough amplitude appearing at the output of the 8,000 c.p.s. filter will operate relay A. This will trigger the multivibrator and relay B de-energises for a time span determined by the time constant of the multi-vibrator. The *b* contacts change into position 1 but relay C will still not operate because relay A is now energised. If the output from the filter ceasea before the multivibrator returns to normal then relay A will release while relay B is atill de-energised and relay C will therefore operate. The contacts of relay C connect a voltage to the maximum detector indicating the recognition of the phoneme /t/. If, however, the output from the 8,000 c.p.s. filter is long in duration and therefore disappears later than the end of the multivibrator cycle, then relay B operates again before relay A releaaes and relay C will not operate at all. This then satisfies the basic condition of obtaining an output from the circuit when the fricative energy ia shorter in duration than a pre-set limiting value and no output when the fricative energy lasts longer than this critical duration. By definition, the choice between a plosive and any other phoneme is not necessarily made until the end of the multivibrator cycle and therefore the rest of the automatic recogniser muat be prevented from making a decision until the end of this period. Thia is achieved by connecting, for the duration of the multivibrator cycle, a voltage greater than the maximum that any multiplier can provide to the input of a special "blind" channel of the maximum detector. The effect of this will be that this special valve, which provides no output, will capture all the current in the maximum detector and theredore the other channels cannot operate and no output can be provided by the maximum detector. The inhibiting voltage is derived from the HT line of the plosive detector. When relay A operatea and triggers the multivibrator, it alao provides the inhibiting voltage through the 1M ohm resistance. If the sound is a short one then a /t/ ia indicated because relay A releases, energiaing relay C, and at the same time removing the inhibiting voltage. If the sound is long, then relay A remain energised but at the end of the multivibrator cycle relay B operates again and its contacts will ahort out the inhibiting voltages and thereby allow the maximum detector to make an alternative choice. The purpose of the resiatance in the grid circuit of V_2 is to discharge the 0.0015 uF condenser rapidly after relay A releases and the diode D_1 ensures that the negative voltage which would otherwise appear at the grid of V_2 during the discharge of the 0.0015 uF condenser is shorted out and does not interfere with the cycle of the multivibrator.

On investigating records similar to those shown in Fig. 10, it was found that the duration of the fricative energy in the 8,000 c.p.s. filter for /t/ was never longer than 50 msec. and for /s/ never shorter than about 200 msec. Other sounds, such as /ʃ/ for

example, also produced energy in the 8,000 c.p.s. filter, but although the duration in these cases was sometimes quite short, the amplitude was considerably smaller than that for the great majority of the /t/ sounds. It was decided therefore to set the threshold of operation of the /t/ detector fairly high; this avoided spurious operation by unwanted signals and also ensured that the /t/ spikes were well within 50 msec. in duration as the width of the spikes was narrower near their peak amplitudes. The duration of the multivibrator cycle was set to 60 msecs.

The circuit for the /k/ detector was identical with that of the /t/ detector but its input was derived from the filter centred on 1,600 c.p.s. As will be seen later when the results are presented, the operation of the /t/ detector was fairly successful. The /k/ detector was very markedly less successful, although quite distinct voltage spikes could be observed for many of the /k/-s. Unfortunately quite a number of other sounds, but particularly /ʃ/, whilst producing a voltage that was noticeably longer in duration than the /k/ spikes, had a markedly spiky envelope which could not be eliminated by smoothing without obliterating the /k/ spikes as well.

THE AUTOMATIC RECOGNITION OF SPACE BETWEEN WORDS

The space between words also had to be treated as a phoneme, as it influenced the statistical distribution of the other phonemes. It was recognised by a simple circuit which derived its input from the main power amplifier. The circuit, shown in Fig. 18, consists of

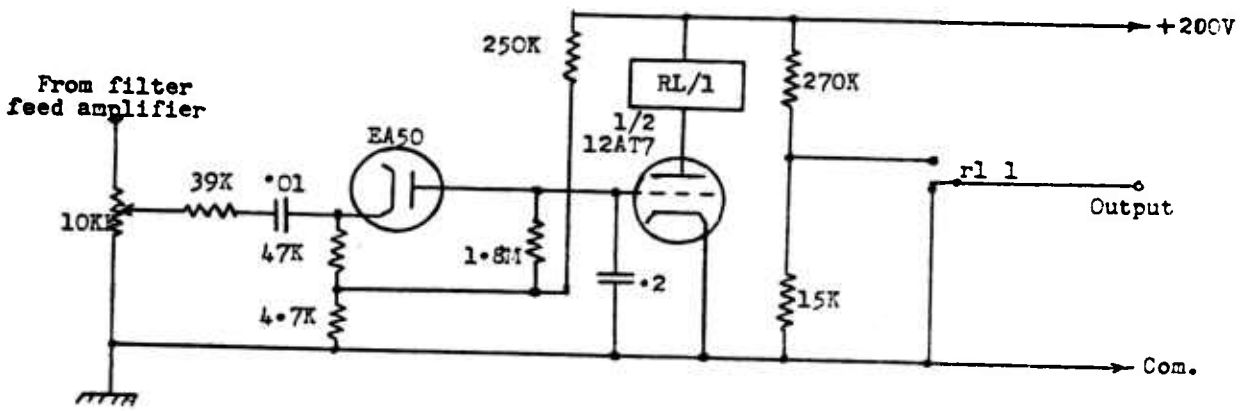


Fig. 18. Simplified circuit diagram of the space detector.

a single triode with a relay in its anode circuit. The valve is normally biased so that the relay is energised and its contacts connect a voltage to one of the inputs of the maximum detector, thus indicating the presence of silence. As soon as a speech voltage appears at the input of the detector, it will be rectified and the rectified voltage, applied to the grid of the triode, will bias it to cut-off and the relay de-energises. The time constant of the rectifier circuit is short for charging and long for decay. This is necessary so that the relay will give a quick indication of the beginning of speech but at the same time does not operate during the short gaps that occur in speech, for example during stop consonants. The circuit will in fact give an indication well under 10 msec. after the beginning of speech but will have a delay of about one second at the

end of words. This is a longer delay than is really required as a break in speech energy of more than 0.25 sec. during words was never observed.

Referring again to Fig. 18, it will be observed that the decay time of the circuit will be affected by the setting of the input potentiometer because this will determine how far the triode is biased off. Also, the positive bias applied to the grid of the triode through the 250 K ohm and 4.7 K ohm resistances will raise the threshold of operation of the circuit so as to avoid spurious operation by noise. A pen recording demonstrating the operation of the space detector is shown in Fig. 19. The upper trace shows the speech input rectified

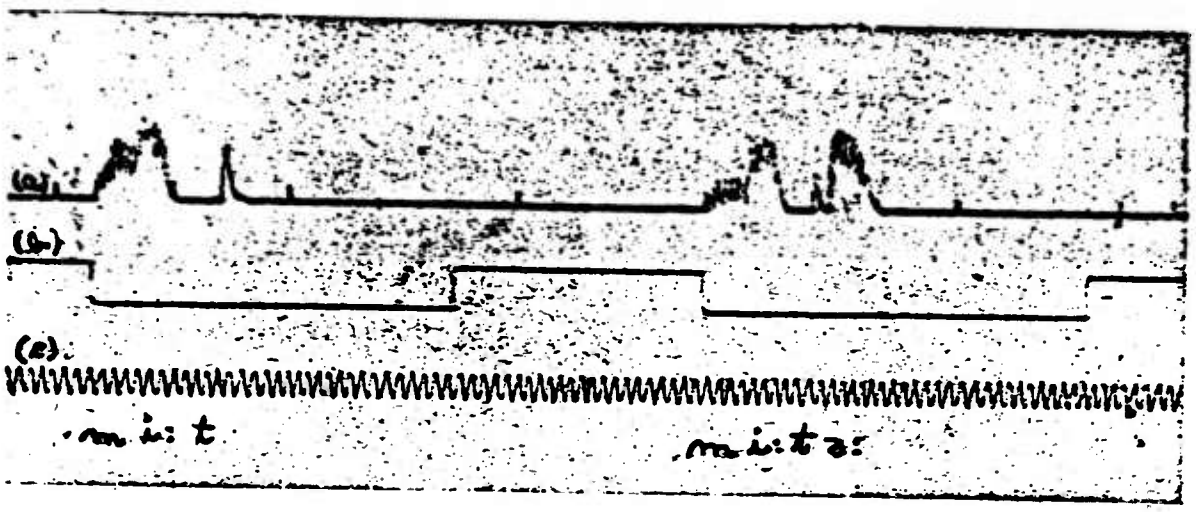


Fig. 19 Pen recording illustrating the action of the space detector.

(a) Speech envelope.
(b) Space detector output.
(c) 20 c/s time marker.

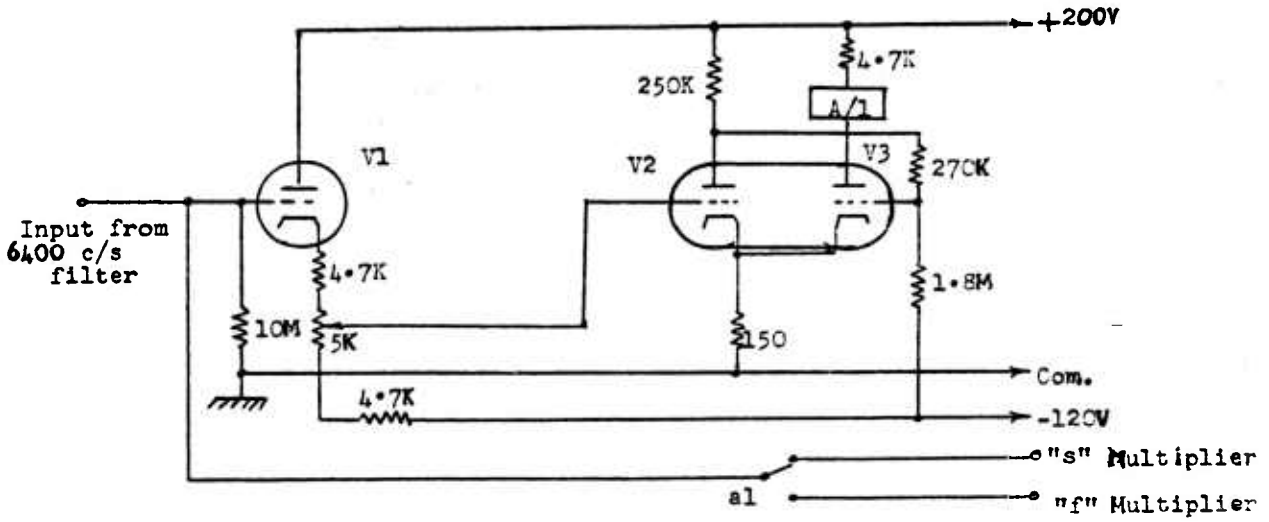
and smoothed with a low pass filter of 10 msec. time constant. The middle trace shows the output of the space detector, the upper position of the pen corresponding to the space indication. The third trace at the bottom was produced by a 20 c.p.s. time marker, and indicates a paper speed of about 1¼ in./sec. The recording shows clearly the quick response to a speech input and the slow decay and also the insensitivity to spurious noise.

This completes the description of the acoustic recognition circuits used in the version of the automatic recogniser with which many of the experiments were carried out. Later on, however, it was attempted to add two more sounds, /f/ and /z/, to the vocabulary of the recogniser and suitable circuits for acoustic recognition had to be found for them. Both of these used not only spectral but also other kinds of cues for recognition.

THE AUTOMATIC RECOGNITION OF /f/ and /z/.

Published work on the analysis and the recognition of fricatives (30) (33) (53) suggested that intensity as well as spectral pattern might help in the recognition of

/f/. The examination of spectral patterns for /f/ sounds by means of the 18 channel filter bank showed that the spectral maximum for /f/ is in about the same region as for /s/ but with a markedly reduced amplitude for /f/ compared with /s/. The usual multiplier circuit was therefore used for recognising /f/: the same spectral pattern was to be detected as for /s/ and therefore the two inputs to the /f/ multiplier circuit were derived from the 6,400 c.p.s. filters just as for the /s/ detector. An amplitude dependent selection between /f/ and /s/ is achieved by not connecting the output of the 6,400 c.p.s. filter to the inputs of the two multipliers directly but by switching it instead to the input of the /f/ multiplier if the output of the filter is below a certain threshold value and to the /s/ multiplier if it is above the threshold voltage; the output of the 8,000 c.p.s. filter is connected permanently to the inputs of both multipliers. In this way only one of these two multipliers can give an output at any one time, the one to which the 6,400 c.p.s. filter is connected. The amplitude dependent switching of the filter output is achieved by the circuit shown in Fig. 20.



V1 - CV133 V2 & V3 - 12AT7

Fig. 20. Diagram of amplitude discriminating circuit for /f/ detector.

The filter output is connected to a cathode follower input stage and from there to a Schmidt trigger circuit formed by triodes V_2 and V_3. The coil of a relay is connected to the anode circuit of V_3 and in the quiescent state the relay is energised and its contacts connect the output of the filter to the input of the /f/ multiplier. When the filter output exceeds a certain value the Schmidt trigger operates, the relay de-energises and the output of the filter is switched to the /s/ multiplier. The total range of voltages from the 6,400 c.p.s. filter is about 65 volts, the change-over point of the trigger is set to about 50 volts and the sensitivity of the trigger is about 1 volt.

As far as the recognition of /z/ is concerned, it was soon realised that the acoustic features of this phoneme do not necessarily consist of the conventional concept of a hiss modulated by vocal cord tone but is often distinguished by being shorter and weaker than its unvoiced counterpart /s/. The best way of differentiating it acoustically from /s/ therefore is probably not by searching for energy concentration in the low frequency part of the spectrum but by looking instead for the acoustic expression of the fortis-lenis opposition. Previous work specifically concerned with the acoustic cues for the /s/ - /z/ distinction (5) has shown that the duration and the intensity of the hiss play a part in making this

distinction. Although the results referred to also indicate that it is the relation of the duration and of the intensity of the hiss and of the preceding vowel sound that serve as a cue for recognition, it was decided to try /z/ recognition by measuring the absolute value of the hiss associated with the speaking of the /z/ phoneme. A similar circuit to that already tried for the recognition of /t/ was used: a relay operated by a one-shot multivibrator switched the output of the 6,400 c.p.s. filter either to the /s/ or to the /z/ multiplier circuit, depending on whether the duration of the hiss was smaller or greater than that of the period of the multivibrator. The duration of the multivibrator cycle was about 130 msec. The duration of the hiss then helped to distinguish between three phonemes: the /s/, /z/ and /t/. If the duration of the friction was 60 msec. or less then a /t/ was recognised, if it was between 60 and 130 msec. a /z/ and if it was more than 130 msec. an /s/ was indicated.

The circuits for acoustic recognition could have been improved in a number of different ways: a better-designed bank of filters, more inputs to the multipliers to allow for more complex spectral reference patterns, the use of a certain amount of memory to detect relative values of the duration and intensity of successive segments and to detect formant changes (for the recognition of transitions), etc. It has already been indicated earlier, however, that there was no intention of pursuing the question of acoustic recognition further than was necessary to achieve reasonable success so that the effects of using linguistic information could be investigated. The acoustic recogniser was therefore not developed beyond the circuits already described and the following paragraphs will deal with the circuits which implement the linguistic part of the recogniser.

THE STORAGE AND USE OF LINGUISTIC INFORMATION

The linguistic information that was to be used in the recognition process consisted of the digram frequencies of the phonemes; at any point in the recognition process therefore information had to be available about the probability of occurrence of the various phonemes in the repertory of the machine as a function of the preceding phoneme. This meant that the digram frequency of all possible phoneme combinations had to be stored in the machine: if there were n phonemes then there would be n^2 digram frequencies to store. As the recognition proceeded phoneme by phoneme the previously recognised phoneme had always to be remembered and in the light of its identity the linguistic store had to provide a different set of n digram frequencies for use in the recognition process. This means that two separate memories were needed for using the linguistic information: a permanent one which remembered all n^2 digram frequencies and a continuously changing one which remembered the identity of a phoneme recognised for the duration of the next phoneme only and then remembered the identity of this fresh phoneme for the duration of the one after that and so on; this second memory determined which set of n digram frequencies appeared at the output of the linguistic store.

The digram frequencies are remembered in the form of the setting of potentiometer sliders: when a fixed voltage is applied across the potentiometer the slider provides a voltage proportional to the digram frequency. The complete store consists of n^2 potentiometers, one for each of the n^2 different digram frequencies to be remembered. The potentiometers are arranged in n columns of n potentiometers each. The ends of the potentiometers in any one column are connected together and the slider of each one is set proportional to the digram of frequency of the n different phonemes following one of the phonemes. When a voltage is applied to the commoned ends of a column of potentiometers, the sliders will provide voltages that are proportional to the digram frequencies. There are n similar columns, one for each phoneme or rather one for each set of digram frequencies following these phonemes. Part of such a matrix of potentiometers is shown in Fig. 21. If, for instance, the previous phoneme recognised has been /m/, then a fixed voltage is applied to the /m/ column and the sliders provide a set of voltages proportional to the

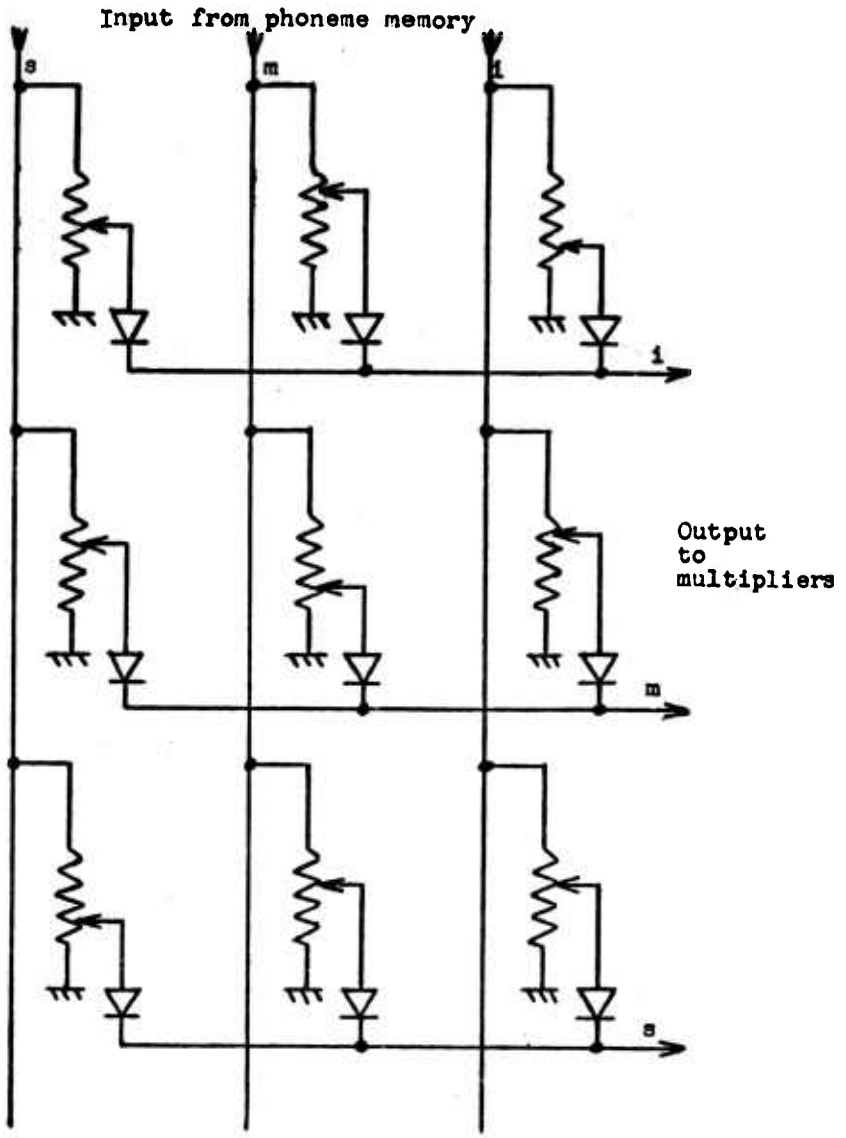


Fig. 21. A typical section of the "store of linguistic knowledge" circuit.

digram frequencies of the various phonemes following /m/. These voltages are then led by the common rails, shown horizontally, to the recogniser circuits and are used in ways to be described later. The diodes, shown in the circuit diagram, prevent the loading of the output voltages by the potentiometers not energised: all these voltages are positive so that current can flow from the potentiometers to the horizontal rails but not in the opposite direction. The actual memory consisted of 256 potentiometers assembled into a 16 by 16 matrix which allowed memory space for remembering the digram frequencies of as many as 16 phonemes, although the full capacity was never used. The resistance of each potentiometer was 10 K Ω and the slider was adjusted by a screwdriver. Two rotary switches and a voltmeter were also provided on the same panel as the potentiometers. One of the switches, marked "input", could permanently switch a fixed D.C. voltage to any column of

potentiometers overriding the control by the maximum detector and relay memory. The other switch, marked "output", could connect the voltmeter to any of the horizontal rails of the potentiometer matrix which are in turn connected to the sliders. This arrangement made it possible to adjust the setting of each slider at leisure and could also be used to operate the recogniser without using linguistic information by permanently connecting the fixed D.C. voltage to one of the columns and setting all the sliders in that column to the same position.

This store of information about digram frequencies has to be operated by another memory which remembers the identity of the previous recognition. It is this "phoneme memory" that provides the voltage which is applied to the appropriate column of potentiometers in the memory of digram frequencies. The phoneme memory is a short-term one and has to operate in conjunction with the maximum detector already mentioned. It must memorise the identity of whichever phoneme is selected by the maximum detector, produce this information as soon as the maximum detector has changed over to another phoneme and then continue to produce this information until the maximum detector makes a new selection. At any one time, therefore, the phoneme memory must be capable of storing information about the phoneme recognition being made currently, but give no corresponding output, and also give an output indicating the identity of the previous phoneme recognition. The changeover, when the existing output is to be discarded to be replaced by information about the phoneme which has been stored, takes place whenever the maximum detector changes from one selection to another.

The circuit developed for this purpose consists of a number of identical units, one unit for each phoneme being dealt with by the recogniser, and each unit is associated with a different one of the relays in the anodes of the maximum detector. The circuit diagram of one of these memory units is shown in Fig. 22. In the quiescent

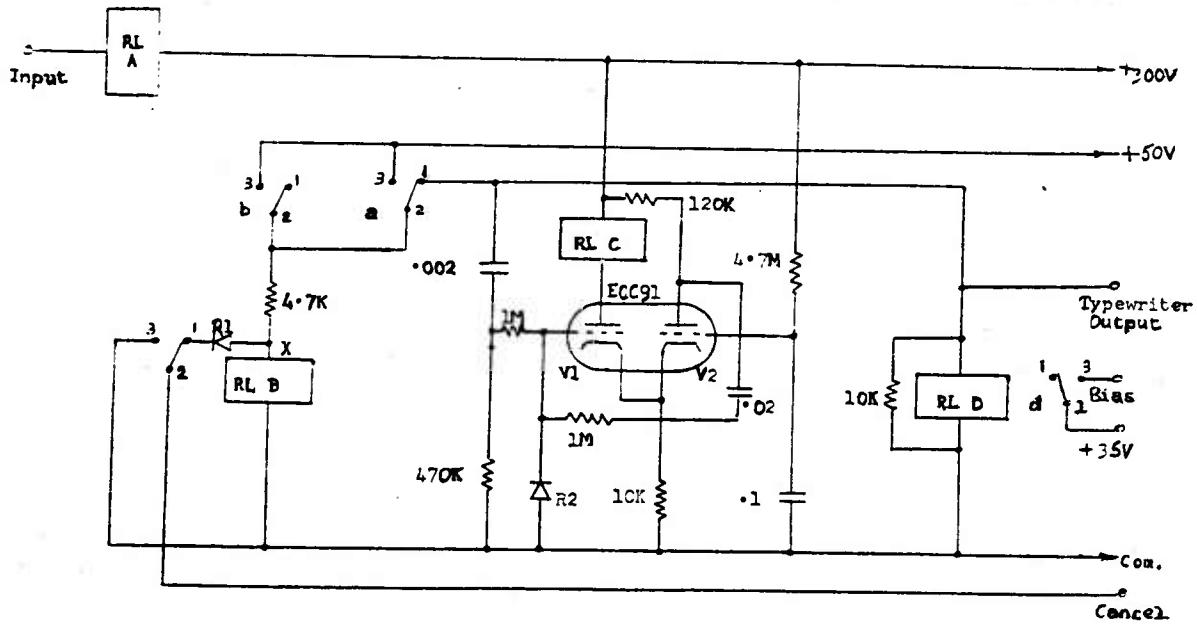


Fig. 22. Circuit diagram of a typical phoneme memory unit.

state all the relays are de-energised. Relay A is the relay of the maximum detector circuit. When it energises, indicating that the maximum detector has selected that particular phoneme, the *a* contacts operate relay B which is self-holding through the *b* contacts. The operation of the B relay constitutes the

memorising action which will maintain itself even after *A* releases. When relay A does release, indicating that a fresh maximum has appeared and has been selected by the operation of another A relay in the maximum detector, then the *a* contacts return to position 1 and the 50 volt H.T. supply is applied to the line marked "typewriter output" through position 3 of the *b* contacts and position 1 of the *a* contacts. This will operate the typewriter and a character is printed identifying the phoneme recognised; this means that the typewriter indicates a recognition at the moment when the actual recogniser has started the next recognition. The 50 volt supply also operates relay D, the contacts of which apply the fixed voltage to the appropriate column of potentiometers in the digram frequency store. Finally the 50 volt supply, on being applied to the "typewriter output" line, will also trigger the one-shot multivibrator circuit of valves V_1 and V_2 and relay C operates for the duration of the cycle of this multivibrator. Position 2 of the *c* contacts of all the memory units are commoned to the line marked "cancel"; although this means that the points X on all the units are connected in parallel, the 50 volts supplied by the *a* or *b* contacts of any one unit will still only affect the B relay of their own unit because of the diode R_1. When the *c* contacts of one of the units change over from position 1 to 3, the "cancel" line is shorted to earth and the B relay of any unit that might have been energised is released, except for the one in which the C relay has operated because in that unit the shorting action of the *c* contacts has at the same time disconnected that B relay from the cancel line. The diode R_1 is biased in the forward direction now and will not impede the shorting action. In this way the memory of all the units is cleared, except the one that has been operated immediately before. The multivibrator and relay C will operate for only a short time when relay A releases; relay D however will remain operated as long as relay B is energised. Relay B remains energised until the A relay in another unit releases. This will operate the C relay of that unit and the shorting action of its contacts will cancel the B relay of the first memory unit. Once the coil of the B relay is shorted its contacts release and the first unit is returned to its quiescent state and is ready for another operation. Summarising briefly, then, the A relay activates the memory unit and the B relay retains the memory without, at this time, giving an output. When the A relay releases, the C relay cancels all previous memories but retains its own, and the D relay activates the relevant part of the memory of digram frequencies.

All the relays in these circuits are of the high-speed type that operate and release in less than 2 msecs. and the speed of operation of the circuit is controlled by the duration of the multivibrator cycle which is purposely made as long as about 35 msec. This seems a long period but it must be remembered that at the rate of speaking used in the speech input to be recognised the average duration of phonemes is about 300 msec. and the duration of all but a few phonemes, plosives for example, is more than 170 msec. Most of the recogniser circuitry is rendered inoperative during the time that any one of the C relays is energised because the *c* contacts short the common "cancel" line to earth and prevent the operation of another B relay during this interval. This period of immobility is needed to prevent undesired operation of the recogniser circuits owing to a variety of reasons. For instance, immediately after the change-over some time is needed, because of relay and multiplier circuit time constants, before the fresh information about digram frequencies is fully effective and the wrong maximum might be selected during this interregnum. Also as is well-known, formant transitions take place at the phoneme boundaries which might give rise to spectral patterns of short duration that match quite different spectral reference patterns from those matched best by the succeeding steady state segment. The nature of these transitions provides, of course, valuable cues for recognition but they are not used in the system of recognition discussed here and therefore it is preferable to exclude their effects. These and other considerations made it necessary

to set the time constant of the multivibrator to 35 msec. and thereby to exclude recognition *by means of pattern matching* of any event with a shorter duration than this period.

The foregoing paragraphs have explained how the information about digram frequencies was stored and made available for use at the right time. The information thus provided was applied to the recognition process by a further stage of multiplication. The multipliers of the acoustic recogniser provide a set of voltages, one voltage for each phoneme, showing how well the input wave corresponds with the phonemic reference patterns or in other words showing the relative likelihood of occurrence of the phonemes from the acoustic point of view. At the same time, another set of voltages is also available from the store of linguistic knowledge, again one voltage for each phoneme, which are an expression of the likelihood of occurrence of the various phonemes from the linguistic point of view. These two streams of information are combined by multiplying separately for each phoneme the acoustically derived voltage with the corresponding voltage from the linguistic store and then selecting the largest product. This means that for each phoneme two multiplications are carried out prior to maximum selection. The two filter outputs are multiplied, as explained previously, for acoustic recognition and the product is then multiplied with the voltage representing the appropriate digram frequency and this second product is applied to the maximum detector. In constructing the recogniser two identical multiplier circuits were mounted on a common sub-chassis to take care of this double multiplication process for one phoneme and there were as many of these double multiplier circuits as there were phonemes to be recognised by pattern matching. A schematic diagram of the arrangement for the complete recogniser is shown in Fig. 23. Each of the boxes marked

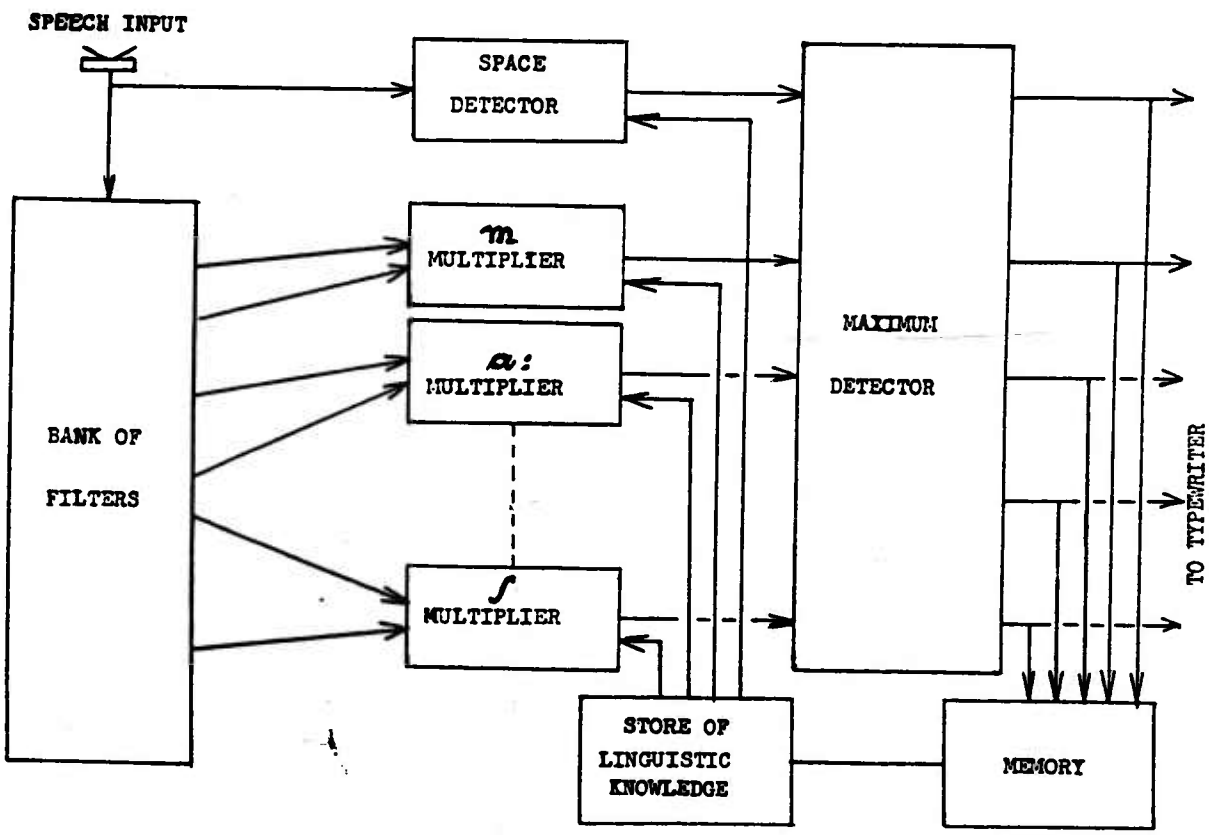


Fig. 23. Schematic diagram showing the arrangement by which acoustic and linguistic information are combined in the recogniser.

'multiplier" multiplies three voltages, two from the filters and one from the 'store of linguistic knowledge'. All the multipliers work simultaneously and the 'maximum detector" identifies that "multiplier" which provides the largest product. The output of the maximum detector operates the phoneme memory and the typewriter (not shown in Fig. 23).

THE OPERATION OF THE TYPEWRITER AND THE TYPEWRITER MEMORY

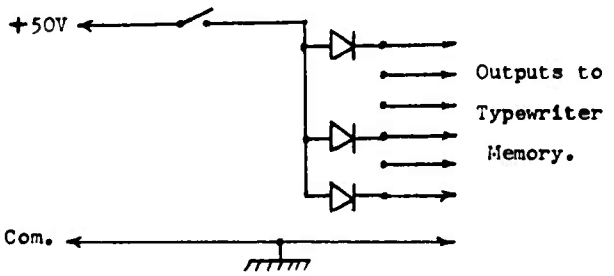
Next, the arrangement by which the appropriate key of the typewriter is operated will be described. An Underwood electric typewriter was used with the recogniser. Unfortunately no electrical typewriter exists in which the keys are operated by electrical action: they are all mechanically operated and are "electrical" only in the sense that the initial movement, which must be produced mechanically, triggers off the movement of the type bar, this latter movement deriving its energy from an electrically driven roller. A set of solenoids had to be added therefore to the typewriter to provide the initial movement of the individual keys. The solenoids were also supplied by Underwoods. It was recommended that the solenoids should be mounted underneath the keyboard and arranged to pull the keys downwards. In the end it was found more convenient to mount the solenoids above the keyboard and arrange that the plungers of the solenoids push the keys downwards. This avoided the need for hooking each plunger to its appropriate key-bar as would have been necessary if the solenoids had been mounted underneath the keyboard. The plungers are returned to their normal position by the return springs of the typewriter keys. These springs had to be reinforced by phosphor-bronze sheet arranged to press the keys upwards. A fresh set of typewriter key tops were fixed to the tops of the solenoid plungers so that the typewriter can be operated by hand as well as by actuating the solenoids.

The solenoids needed a fairly high current if they were to operate sufficiently fast; at the same time the total energy consumption had to be limited to avoid overheating. Both requirements could be satisfied by operating the solenoids with a measured amount of energy from a condenser charged to a high voltage. The contacts of the maximum detector relays connected a 64 uF condenser charged to 300 volts across the solenoid; when the relay de-energised the condenser was recharged by being connected to a 300 volt D.C. line. The amount of energy stored in the 64 uF condenser was sufficient to move the plunger of the solenoids and to operate the typewriter. The resistance of the solenoids was 100 ohms and the peak current flowing was therefore around 3 amps. Unfortunately this was too high for the contacts of the high speed relays that are used in the maximum detector; no relay with contacts having a sufficiently high current rating and at the same time operating at the high speeds required could be found and therefore the solenoids were operated by a suitably triggered thyratron. The thyratron circuit will be described later, after another modification of the circuitry for operating the typewriter has been discussed.

The maximum rate at which the typewriter can type is about 15 characters per second. The average rate at which phonemes succeed each other in the speech material used for testing the recogniser was only about one per second if the interval between words is taken as part of the speech and still only about 9 per second during the words themselves. Despite these low average rates, the peak rate was considerably higher and when watching the typewriter during the operation of the recogniser it became clear that it failed to

type some of the phonemes selected by the recogniser: on some occasions the solenoid plungers moved but not the type, on other occasions several type bars operated in quick succession, became entangled and jammed the typewriter. The peak rate at which phonemes were being recognised, which was later found to be about 20 to 25 per second, was therefore too high for the typewriter, whilst the average rate was well within its capabilities. It was thought possible therefore to overcome this difficulty by using an "information rate smoother", a device which accepts and stores the phoneme recognitions at whatever rate they occur and then "reads" them into the typewriter at a constant, slower rate. This rate must be less than the maximum of which the typewriter is capable and as long as the output rate is higher than the average rate of phoneme recognition the system will work satisfactorily with quite modest storage requirements.

A storage system based on magnetic tape was first tried but was not a success because of the difficulty of controlling the movement of the tape in a satisfactory manner. Next a system in which a bank of condensers was used as the information store was tried and was found to be satisfactory. In this system, each phoneme was designated by a binary number and as all 44 keys of the typewriter were to be catered for a six-digit binary code was used. Whenever a phoneme was recognised the corresponding binary number was produced by means of a simple diode coding network shown in Fig. 24. A diode



Note: The outputs of all panels are connected in parallel.

Fig. 24. Example of coding network.
The 50 volt supply is obtained from the "typewriter output" line of the phoneme memory (fig. 22.) and the single switch on this diagram represents the combination of the contacts of relay A (off) and relay B (on) of fig. 22.

was connected in whichever position a binary one was needed and the diode was omitted for a binary zero. There were, of course, as many different coding panels as there were symbols to be typed. The binary number corresponding to each phoneme recognition was stored in the form of unit charges on a number of condensers. As a six-digit code was used, six condensers were needed to store one number. Altogether 25 rows of 6 condensers each were provided so that 25 successive binary numbers could be stored. One end of all condensers was connected together and earthed. The other ends were connected to two 25 position multi-level uniselectors in such a way that any one condenser was

connected to the two corresponding positions of the two uniselectors. The six condensers storing one binary number were connected to six different levels of the same position on the uniselectors, the next six condensers to the six levels of the next position, and so on. The six wipers of the first uniselector, called the "write" uniselector, receive the six voltages (or lack of voltage) representing the six digits of the binary number to be stored and charge the six condensers, in the position to which the wipers are connected, accordingly. The wipers then step to the next position and the next binary number is stored and so on. The "write" uniselector is stepped by the cancelling pulse of the phoneme memory (Fig.22) so that it steps every time a new recognition is made. The stepping of the wipers of the second uniselector, called the "read" uniselector, is determined by the frequency of a multivibrator provided for this purpose; this frequency was adjustable and was set to approximately 2 c.p.s. A seventh level on both write and read uniselectors was used to ensure that the wipers of the read uniselector could never pass those of the write uniselector. Whenever the "read" wipers get within 2 steps behind the "write" wipers the interlock prevents any further movement of the "read" wipers until the "write" switch has started moving again. A simplified form of the circuit described so far is shown in Fig. 25. It will be seen

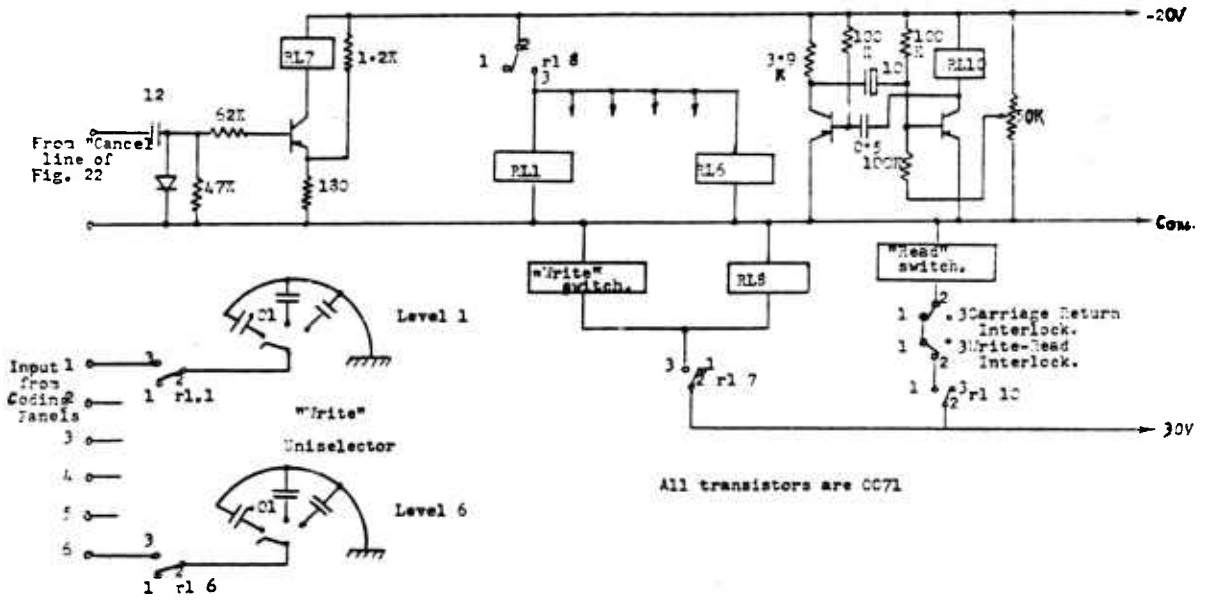


Fig. 25. Circuit diagram of "write" and "read" uniselector switch connections.

that transistors are used rather than valves. This achieved the well-known savings in space required, heat dissipated, wiring needed and in the size and complexity of the power pack; the rest of the recogniser circuitry used valves because transistors were not yet freely available at the time of its design. In the circuit of Fig. 25, the negative-going pulse on the "cancel" line of the phoneme memory is used to drive the "write" uniselector. The pulse drives relay 7, the contacts of which energise the coils of the uniselector switch and of relay 8, which in turn operates relays 1 to 6. The pulse of the "cancel" line lasts 35 msec. and appears every time the phoneme memory indicates a fresh phoneme recognition. The six voltages representing the binary number which stands for the phoneme just recognised appear simultaneously with the beginning of the "cancel" pulse. They will charge the 6 condensers through the contacts of relays 1 to 6 and the wipers of the uniselector. The time constant

of the condenser charge and discharge circuits is only a few microseconds, so that any previous charge on any of the condensers will disappear and the new charges will establish themselves in a relatively very short time. At the end of the 35 msec. "cancel" pulse, relay 7 releases and this will de-energise the coils of relay 8 and of the uniselector. The de-energising of the uniselector coil will make the wipers move forward by one step and the release of relay 8 will open the contacts RL1 to RL6, isolating the condensers which will hold their charge sufficiently for a matter of minutes. The circuit driving the coil of the "read" uniselector is shown at the right-hand end of Fig. 25. A simple, free-running multivibrator is used. It oscillates at about 2 c.p.s. and its frequency can be adjusted by the 50 K ohm potentiometer. The contacts of relay 10 will open and close at the frequency of the multi-vibrator and step the wipers of the uniselector at this rate. The contacts ensuring that the "read" switch can never catch up on the "write" switch, as explained earlier, are also shown, as well as another set of contacts which prevent the "read" uniselector from operating the typewriter keys during the carriage return period and make it stop and wait.

The output voltages from the wipers of the "read" uniselector are decoded, that is the correct typewriter solenoid to be operated is selected, by routing the solenoid operating current along the branches of a relay tree whose settings are determined by the condenser charges. A simplified circuit diagram is shown in Fig. 26. The wipers

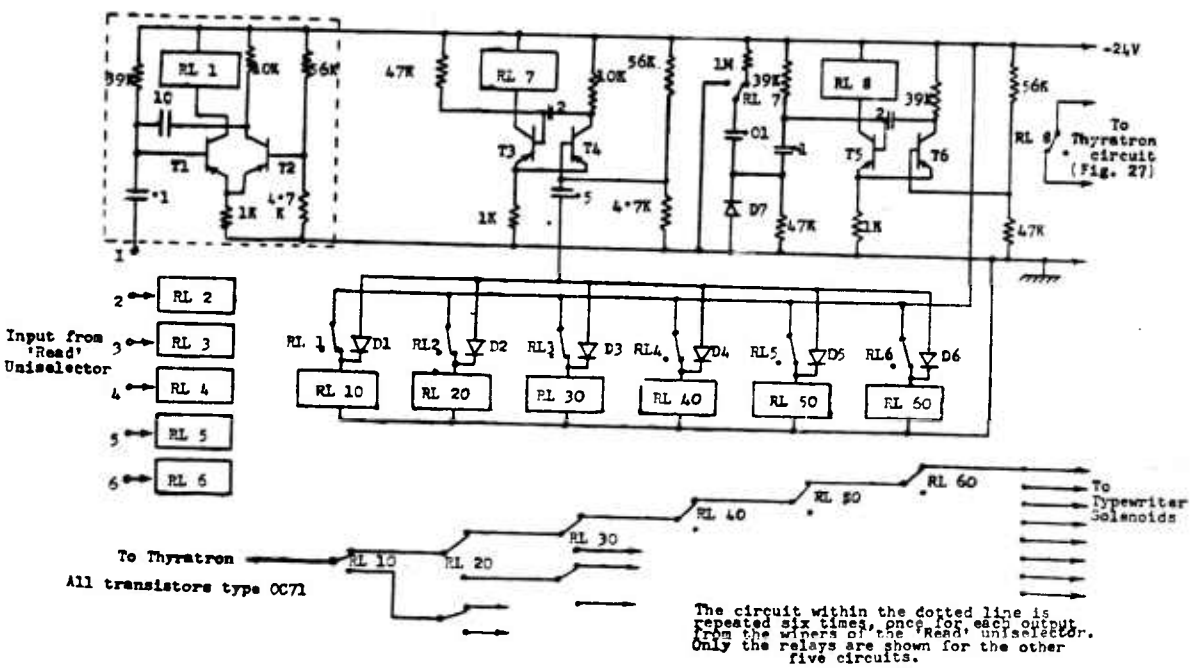


Fig. 25. Simplified diagram of circuit for operating typewriter solenoids from binary coded control input voltage.

of the "read" uni-selector are connected to terminals 1 to 6 on the left. A voltage from, for example, wiper No. 1 will trigger the flip-flop circuit of transistors T_1 and T_2 and will de-energise relay 1. The same happens to any of the other five similar flip-flop circuits, if the wipers to which they are connected carry a binary "one" voltage. The flip-flops remain quiescent if the wipers carry a binary "zero" voltage. The operation of relays 1 to 6 energises relays 10 to 60 and the contacts of these latter relays form the binary decision tree that connects the typewriter solenoids to

the thyratron which provides the energy for operating the solenoids. Relays 10 to 60, like all other relays in the circuit, are of the high-speed variety which carry only one set of change-over contacts so that although only a single relay is shown, in fact a separate relay has to be provided for each set of contacts. Summarizing the operation of the circuit described so far, the 6 digit binary voltage from the uniselector triggers the flip-flops, which operate relays 1 to 6 which in turn operate relays 10 to 60 and the contacts of these pre-select one particular solenoid to which the current to be supplied by the thyratron will be sent. The thyratron circuit, to be described later, is triggered by the two flip-flops of transistors T_3 T_4 T_5 T_6. The operation of relays 1 to 6 will not only prepare the relay tree but also send a triggering pulse through one or more of diodes D_1 to D_6 to the flip-flop of T_3 and T_4. Relay 7 in the collector circuit of T_3 is normally operated and will release for the duration of the cycle of the flip-flop. The contacts of relay 7 normally short the 0.01 uF condenser in the base circuit of T_5 to earth. When relay 7 operates, the condenser will charge to a negative voltage but will not affect T_5 because of the shorting diode D_7. When relay 7 operates again at the end of the cycle of flip-flop T_3 T_4, the condenser is discharged to earth and sets up a positive pulse across the 47 K ohm resistance in the base circuit of T_5; this triggers the flip-flop of transistors T_5 and T_6. Relay 8 in the collector circuit of T_5 is normally operated and will release for the duration of the duty cycle of the flip-flop T_5 T_6. This, as will be seen below, triggers the thyratron and an energising current is sent along the relay tree to the appropriate typewriter solenoid. The flip-flop T_3 T_4 de-energises relay 7 and delays the triggering of flip-flop T_5 T_6 for a time span of 20 msecs. The double purpose of this interval is to allow plenty of time for the contacts of relays 10 to 60 to select the desired path for the solenoid operating current before relay 8 triggers the thyratron and to ensure that the relay contacts do not themselves switch the solenoid current. The contacts of relay 8 remain open for 30 msec. as required by the thyratron circuit and, as will be explained later, the current from the thyratron will cease not later than a further 10 msec. The flip-flop T_1 T_2, which initiated the cycle, keeps relays 1 to 6 and 10 to 60 operated for 200 msec. and therefore at least 140 msec. must elapse after the solenoid current ceases before the contacts of the relay tree change again. This means that the contacts of the relay tree do not either break or make the relatively high solenoid operating current; this feature of the design is important to prevent early damage to the relay contacts. At the end of the 200 msec. operating period of relays 1 to 6 the circuit is ready to receive the next input, derived from the wipers of the "read" uniselector after it has stepped again.

The current for operating the solenoids is switched on and off by a suitably triggered thyratron. As already mentioned, the solenoids need an operating current of about 2 amps but this value of current must not be maintained for longer than about 10 msec., to prevent over-heating. It was not found easy to obtain a mechanical switch that combined the fast operation with the necessary current-carrying capacity and that is why a thyratron was used: the thyratron was triggered to connect the solenoids to the mains supply for one half-cycle of the mains voltage. The necessary circuit is shown in Fig. 27. The solenoid selected by the relay tree is connected to the mains voltage through the thyratron; normally the solenoid is isolated but when the thyratron fires the whole of the mains voltage, less the relatively small thyratron maintaining voltage, is connected across the solenoid and a peak current of 2 to 3 amps will be obtained. The grid voltage of the thyratron is obtained by combining a D.C. voltage with a 50 c.p.s. A.C. voltage. The phase of the A.C. grid voltage leads the anode voltage by about 22^{o}; the phase shift is obtained by the 0.025 uF condenser and 300 K ohm resistance. As a result of this phase shift, the grid voltage will reach its most positive value before the middle of the positive half cycle of the anode voltage. The D.C. voltage, with negative polarity of course, is obtained from

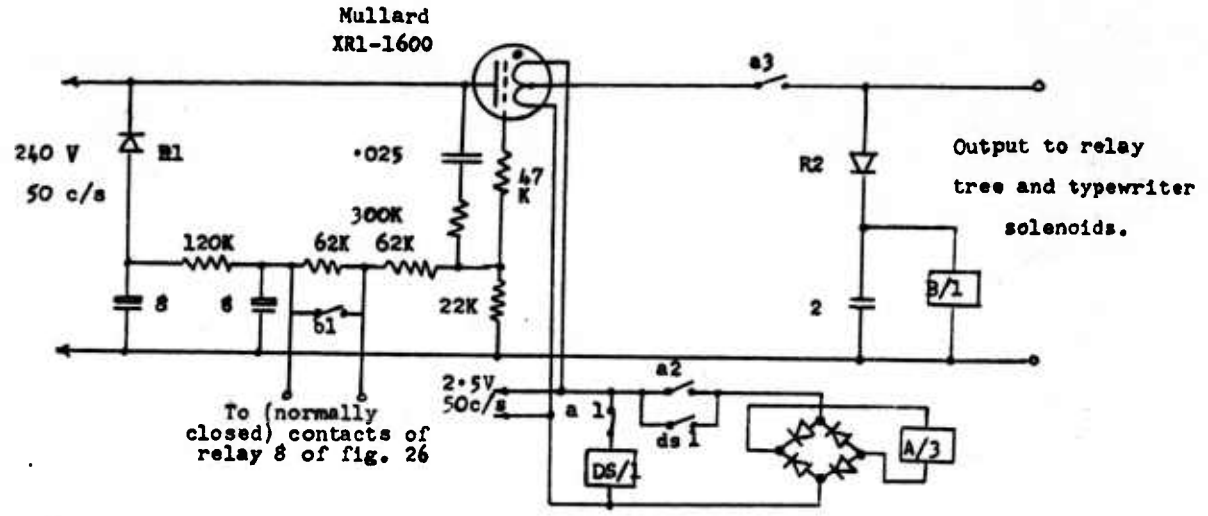


Fig. 27. Circuit diagram of thyratron unit for energising the typewriter solenoids.

the rectifier R_1 and the 8 uF reservoir condenser. In the normal, inoperative condition, the contacts of relay 8 in Fig. 26 short out one of the 62 K ohm resistors, as shown in Fig. 27. Under these conditions the D.C. voltage at the grid of the thyratron is about 33 volts with an A.C. voltage of 15 volts peak value added to it. The triggering voltage of the thyratron, for the peak anode voltage, is around 5 volts and therefore the thyratron cannot strike. When relay 8 of Fig. 26 releases and remains open for 30 msecs., the negative D.C. voltage at the grid falls to about 20 volts and the instantaneous voltage at the grid will fall to around -5 volts during the positive peaks of its A.C. components. As a result the thyratron will fire when the anode voltage is near its peak value. The exact conditions of firing will depend on the instant of operation of relay 8 (Fig. 26) in the A.C. cycle. Two things must be remembered in this connection. First, the duration for which the contacts of relay 8 are open is 30 msec., that is 1½ cycle of the A.C. voltage and second, the thyratron extinguishes during the negative half-cycle of the anode voltage and cannot fire in two successive positive half-cycles of the anode voltage because of the action of relay B. Relay B will operate as soon as the thyratron fires but will not release for about 30 or 40 msec., because the 2 uF condenser can charge rapidly through the rectifier R_2 but must discharge slowly through the high resistance coil of relay B. The contacts of B will short the 62 K ohm resistance and raise the negative grid voltage again. This will not affect the thyratron if it has already fired but will prevent a second firing during the next positive half-cycle, even if the contacts of relay 8 (Fig. 26) are still open. During normal operation, then, if the contacts of relay 8 open during the first half of the positive half-cycle of the anode voltage then the thyratron will fire, then extinguish at the end of the positive half-cycle and will not fire again during the next positive half-cycle, even though the contacts of relay 8, which remain open for 1½ cycles, will probably still be open. If the contacts of relay 8 open during the second half of the positive half-cycle of the anode voltage, the thyratron will not fire because the A.C. component will have made the grid voltage too negative. The thyratron will then fire during the second positive half-cycle of the anode voltage because the contacts of relay 8 remain open for 30 msec. If the contacts of relay 8 first close during the negative half-cycle of the anode voltage,

then the thyratron will fire during the next positive half-cycle. In this way the solenoids receive a current pulse which is no shorter than a ¼ cycle and no longer than ½ cycle, whatever the instant at which the contacts of relay 8 first open. The circuit in the lower part of the diagram shows the operation of the thermal delay switch DS which ensures that current cannot be sent through the thyratron until its cathode is fully heated. The function of self-holding relay A is to disconnect the thermal delay switch as soon as it has operated. This ensures that the thyratron cannot be switched on again without awaiting the full delay period if the supply voltage is switched on again soon after it has been switched off.

THE POWER PACKS

The only parts of the circuitry that have not yet been described are the power packs. The current and voltage requirements, other than 6.3 volt 50 c.p.s. heater supplies, are set out in Table 1. As a general rule, stabilised H.T. supplies were employed throughout. They were preferred even if voltage stability was not of primary importance because the stabilisers provided low hum level and low output impedance. The low output impedance made it possible to supply a number of different circuits from the same H.T. supply without danger of undesirable coupling through the output impedance of the power pack. The only H.T. supplies which were left unstabilised were the +275 volt supply for the filter amplifier, the -700 volt supply for the maximum detector (which was stabilised with a neon tube in the maximum detector circuit itself) and the 30 volt supply for energising the uniselector magnets.

The power pack for the filter amplifier consisted of the conventional vacuum diode full-wave rectifier, using a reservoir condenser followed by a single inductance - capacitance smoothing stage. The negative supply for the maximum detector consisted of a vacuum diode half-wave rectifier and a reservoir condenser. The uniselector supply used a selenium diode bridge rectifier and a reservoir condenser.

The stabilised supplies all used the usual full wave rectifier - reservoir condenser arrangement to provide the H.T. voltage. The stabilisers were of three different kinds, depending on the voltage required. The 300 volt and 200 volt stabilisers used the conventional series valve circuit shown in Fig. 28. Two 12E1 valves are used

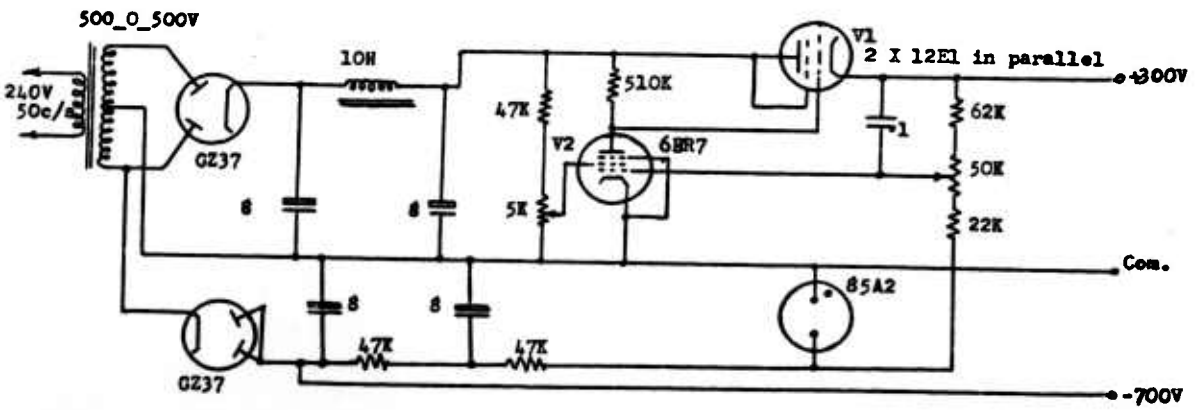


Fig. 28. Simplified circuit diagram of 300 volt series stabiliser.

Table 1. **Voltage and current requirements of the recogniser circuits**

	+300 V	Unstabilised +275 V	+200 V	+50 V	+36 V	−30 V	−120 V	Unstabilised −700 V
Pre-amplifier and filter power amplifier		100 mA						
Filter detector	10 mA							
Multipliers			160 mA				80 mA	
Triangular voltage generator			36 mA					
Plosive detector			120 mA					
f − s discriminator			20 mA				10 mA	
Space detector			20 mA					
Maximum detector	60 mA							35 mA
Phoneme memory	40 mA			50 mA				
Typewriter memory: (1) Storage circuits						100 mA		
(2) Uniselector drive						2 A		
(3) Decoder						300 mA		
Digram frequency store					60 mA			
Total currents	110 mA	100 mA	356 mA	50 mA	60 mA	400 mA and 2 A	90 mA	35 mA

in parallel to serve as the series valve V_1; this doubles the effective mutual conductance and improves the performance of the circuit. The neon stabiliser which provides the reference voltage is supplied from a separate source so that its output is not affected by variations within the stabiliser and as a result the output impedance of the circuit is as low as 0.1 ohm or less, from 1 c.p.s. to about 4 kc.p.s. and is still only 0.3 ohm at 10 kc.p.s. The output impedance rises towards the higher frequencies because the gain of V_2 declines. The mains ripple on the output is about 250 microvolts peak. The output voltage remains constant to about ±½% for a ±10% mains voltage variation. The rectifier cathodes are of the indirectly heated type to ensure that all the valve cathodes in the circuit are fully heated by the time the rectified voltage appears. This prevents the build-up of excessive voltages across some of the valve electrodes and across the smoothing condensers. The adjustable screen grid voltage of V_1 helps in setting the output impedance and hum to a minimum. The circuit can supply up to 200mA at 300 V, although only just over 100 mA were actually required.

The same type of stabiliser was used for the 200 volt supply. Almost 400 mA were required at this voltage and therefore two separate stabilisers, supplying about 200 mA each were used.

The circuit of Fig. 28 is not very suitable for providing low output voltages. The minimum output voltage cannot be less than the sum of the anode - cathode voltage required to operate V_2 and of the grid-cathode voltage required for V_1. In practice the minimum voltage that can conveniently be provided by the circuit is about 200 volts and a different stabiliser was used for the other, lower supply voltages also needed for operating the recogniser. The stabiliser used for the 50 volt supply is shown in Fig. 29; the circuits for obtaining the other low voltages were the same

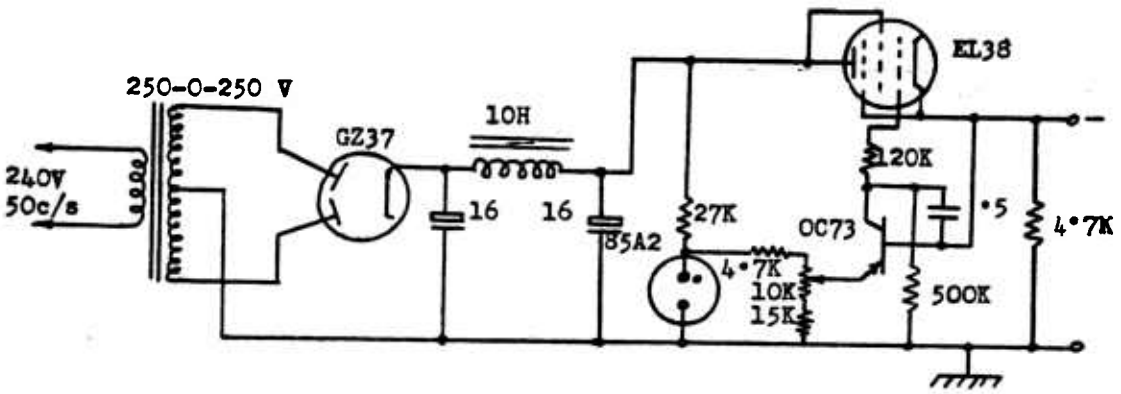


Fig. 29. Simplified circuit diagram of series stabiliser using transistor and valve.

except that a few component values had to be changed to allow for the different output voltages and that the positive output terminal was earthed when an output which is negative relative to earth was required. The circuit of Fig. 29 follows very

closely that described by R.E. Reynolds (46) and is very similar to that of Fig. 28 except that a transistor is used in the amplifier stage instead of a vacuum valve. The transistor amplifier can provide a gain of several hundred and still only requires quite a small voltage across its 500 K ohm load resistance. As a result, the lowest output voltage obtainable from the circuit is about 35 volts which is the grid bias required by the series valve. The 4.7 ohm dummy load resistance ensures that the stabiliser is never used under conditions of no load: the larger grid bias required by the series valve for cut-off would raise the value of the minimum output voltage. The 0.5 uF condenser between base and collector of the transistor prevents self-oscillation. The output impedance of the circuit is about 3 ohms and the output voltage varies by ±0.7% for a ±10% change of the supply voltage.

At the time when the stabiliser circuits shown in Fig. 29 were designed, it was necessary to use vacuum valves in the position of the series valve because no transistor capable of carrying more than about 10 mA was freely available. Soon afterwards, however, the situation changed and it became possible to design series stabilisers in which the bulky series valves which dissipated a large amount of heat could be replaced by a transistor. It was decided therefore to replace the existing power pack for the 30 volt supply, which had to provide 400 mA, with an all-transistor stabiliser. The circuit which was adopted was based on designs given by Brown and Stephenson (2) and is shown in Fig. 30. The transistors T_1 and T_2 are connected in a

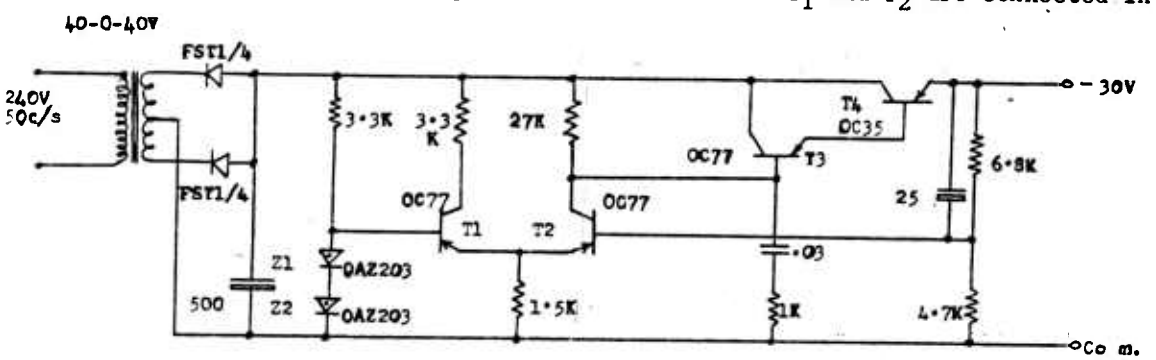


Fig. 30. Circuit diagram of all-transistor series stabiliser.

long-tail pair circuit in which the reference voltage derived from the Zener diodes Z_1 and Z_2 are compared with the output voltage. The amplified voltage difference controls the series transistor T_4. The output of this stabiliser varies by ±¼% for a mains voltage change of ±10%, the output impedance is 0.7 ohms and the ripple on full load is 2 m volts peak.

A photograph of the recogniser can be seen in Fig. 31. The rack which carries the filters is on the left, the multipliers and the maximum detector are on the middle rack. The potentiometer matrix of the digram frequency memory is at the top of the right hand rack; the phoneme memory and the typewriter memory are at the bottom of this rack. The tape recorder which provides the speech input and the typewriter with the solenoids can be seen on the table in front.

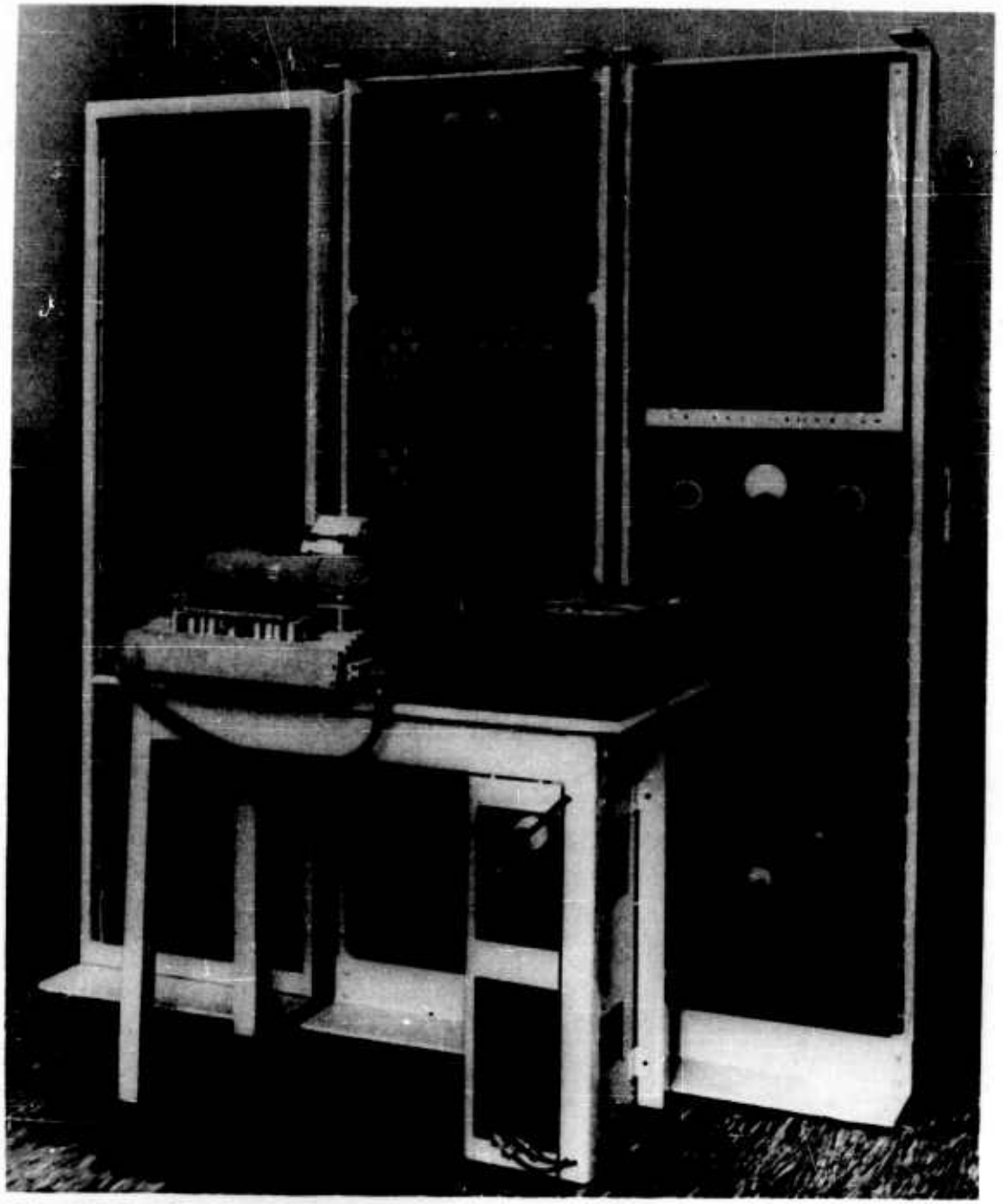


Fig. 31. The automatic phoneme recogniser.

CHAPTER V

THE SPEECH MATERIAL USED FOR TESTING THE RECOGNISER

Having assembled and tested the circuitry that has just been described, suitsble speech material was needed to measure the overall performance of the automatic recogniser. The speech material to be used was, as has already been mentioned, in the form of a list of words which was recorded on magnetic tape for repeated use with the recogniser. The words were spoken in isolation, with an intonation appropriate for a simple statement and the speaker was asked to speak at a constant level as judged subjectively by himself. The same word lists were recorded by three different speakers, although most of the results quoted below were obtsined by using the words spoken by one speaker only.

Several different word lists have been used in the course of the experiments. The first to be used consisted of 139 words; the words contained only the phonemes in the repertory of the machine, that is, /t k s ʃ m n l i: a: u: ə:/, although in the case of the vowels the short /i/ and /ə/ phonemes were also allowed and were put into the same class as the long /i:/ and /ə:/ respectively. In this way, for example, the final vowel in the word *mercy* was considered to be in the same phonemic class as the vowel in *teak* and again both vowels in the word *murmur* were considered to be the same. The word list, both spelt and in the appropriate phonetic transcription is given in Table 2, and will be referred to as list 1. The 139 words

Table 2

List of 139 words used for testing the automatic recogniser

List 1

tart	ta:t	teat	ti:t	teak	ti:k
team	ti:m	teal	ti:l	toot	tu:t
tomb	tu:m	tool	tu:l	Turk	tə:k
terse	tə:s	term	tə:m	cart	ka:t
calm	ka:m	keel	ki:l	keen	ki:n
coot	ku:t	combe	ku:m	coon	ku:n
curt	kə:t	curse	kə:s	curl	kə:l
seat	si:t	seek	si:k	cease	si:s
seem	si:m	seen	si:n	suit	su:t
shark	ʃa:k	sheet	ʃi:t	chic	ʃi:k
shoot	ʃu:t	shirt	ʃə:t	shirk	ʃə:k
mart	ma:t	marsh	ma:ʃ	meet	mi:t
meek	mi:k	meal	mi:l	mean	mi:n

Table 2 (cont.)

moot	mu:t	moose	mu:s	moon	mu:n
murk	mə:k	lark	la:k	leet	li:t
leak	li:k	lease	li:s	leash	li:ʃ
lean	li:n	loot	lu:t	loose	lu:s
loom	lu:m	lurk	lə:k	loon	lu:n
learn	lə:n	tarn	ta:n	turn	tə:n
soon	su:n	sheen	ʃi:n	tartar	ta:tə:
tartan	ta:tn	teeter	ti:tə:	tee-shirt	ti:ʃə:t
Tina	ti:nə:	turkey	tə:ki:	cartoon	ka:tu:n
Khartoum	ka:tu:m	carter	ka:te:	carton	ka:tn
car-seat	ka:si:t	car-mart	ka:ma:t	calmer	ka:mə:
keener	ki:nə:	cooler	ku:lə:	cocoon	kə:ku:n
curser	kə:sə:	curly	kə:li:	curler	kə:lə:
canoe	kə:nu:	seeker	si:kə:	sea-mark	si:ma:k
sooner	su:nə:	sateen	sə:ti:n	Circe	sə:si:
surcease	sə:si:s	Sassoon	sə:su:n	salaam	sə:la:m
salute	sə:lu:t	saloon	sə:lu:n	chenille	ʃə:ni:l
shooter	ʃu:tə:	chi-chi	ʃi:ʃi:	shirker	ʃə:kə:
martyr	ma:tə:	market	ma:kə:t	Marsham	ma:ʃm
Marshall	ma:ʃl	Marner	ma:nə:	meter	mi:tə:
meeker	mi:kə:	meaner	mi:nə:	mercy	mə:si:
mercer	mə:sə:	machine	mə:ʃi:n	murmur	mə:mə:
lama	la:mə:	litre	li:tə:	Lima	li:mə:
leaner	li:nə:	Luton	lu:tn	lucre	lu:kə:
looser	lu:sə:	loosen	lu:sn	Lucerne	lu:sə:n
Lulu	lu:lu:	lunar	lu:nə:	lateen	lə:ti:n
lurker	lə:kə:	learner	lə:nə:	curtain	kə:tn
turtle	tə:tl	certain	sə:tn	castle	ka:sl
Carson	ka:sn	carnal	ka:nl	kirtle	kə:tl
colonel	kə:nl	Seaton	si:tn	seamen	si:mə:n
circle	sə:kl	sermon	sə:mə:n	Martian	ma:ʃn
Meikle	mi:kl	Merton	mə:tn	myrtle	mə:tl
two-seater	tu:si:tə:	Saluki	sə:lu:ki:	lacuna	lə:ku:nə:
merciless	mə:si:lə:s				

Number of words:	139
Number of sounds:	526
Number of digrams:	665
Number of monosyllabic words:	59
Number of disyllabic words:	76
Number of trisyllabic words:	4

are made up of 59 monosyllabic words, 76 disyllabic words and 4 trisyllabic words; the total number of phonemes in the list is 526 and there are 665 digrams, that is phoneme transitions, including the transitions from inter-word space to initial phoneme and from final phoneme to inter-word space. The frequency of occurrence of the phonemes in this list is given in Table 3 and the digram frequencies, again as found in this list, are given in Table 4(a).

Table 3

Frequency of occurrence of phonemes in List 1, expressed as percentages of the total number of phonemes.

t	12%	l	7%
k	10%	ɹ	3%
s	8%	i:	10%
ʃ	4%	a:	6%
m	9%	u:	7%
n	8%	ə:	16%

Table 4(a)

Digram frequencies of phonemes in List 1, expressed as a percentage of the occurrence of each phoneme.

First phoneme \ Second phoneme	t	k	s	ʃ	m	n	l	ɬ	i:	a:	u:	ə:	#
t						11		5	14	6	9	22	33
k								4	9	22	11	31	22
s						4		2	32	2	9	33	18
ʃ					5	5		5	32	5	11	26	11
m									18	22	7	31	22
n								4	2		2	18	74
l									25	8	42	25	
ɬ													100
i:	23	15	6	6	10	23	2	8					8
a:	37	13	10	13	17	10							
u:	22	5	16		11	30	5	3					
ə:	15	12	12	1	4	11	7	1					37
#	15	22	17	8	19		19						

Table 4(b)

Voltage settings of potentiometers in store of linguistic knowledge to represent digram frequencies in Table 4(a).

First phoneme \ Second phoneme	t	k	s	ʃ	m	l	ɬ	n	a:	i:	u:	ə:
t							8	18	10	23	15	36
k									25	11	13	36
s								5		34	10	36
ʃ					6		6	6	6	36	12	30
m									26	21	8	36
l									7	22	36	22
ɬ												
n							9			5	5	36
a:	36	13	10	13	16			10				
i:	36	24	9	9	15		12	36				
u:	26	7	20		13	7		36				
ə:	36	30	30		9	18		27				
#	25	36	28	13	32	32						

List 1 contains quite a number of proper names like *Seaton* or *Carson* and some unusual words like *lacuna* or *chenille*. This was undesirable in some experiments in which a human reader or listener was asked to interpret the output of the recogniser and therefore a second list was made up, to be called List 2, which consists of a selection of the words from List 1, all proper names and many of the unusual words of List 1 having been omitted. List 2 consisted of 75 words and these are given, in the randomised order in which they were used in some of the experiments to be described later, in Table 5. The words were made up of 41 monosyllabic words, 33 disyllabic

Table 5

List of 75 words used for acoustic and for visual presentation of the recogniser output

List 2

cart	seaman	tool	carton
curly	moon	sermon	car-seat
saloon	teat	tomb	meter
leak	shirt	team	seeker
cooler	curtain	keen	tartan
mercy	seek	learn	castle
tart	shooter	turkey	meek
leash	marsh	loose	loot
lease	loosen	term	curt
seat	martyr	colonel	seem
meal	meeker	salute	cartoon
sheen	suit	Turk	certain
sooner	learner	curse	teak
tartar	calmer	cease	market
machine	keener	circle	mean
meaner	soon	seen	carter
loom	shark	sheet	lean
turn	calm	canoe	merciless
lurk	meet	terse	

No. of words:	75	No. of monosyllabic words:	41
No. of sounds:	270	No. of disyllabic words:	33
No. of digrams:	345	No. of trisyllabic words:	1

words and one trisyllabic word. The 75 words contained altogether 270 phonemes and 345 digrams. List 2 was never recorded separately; instead, the recording of List 1 was used as input and the responses of the recogniser to the words of List 2 were selected from the output.

At a later stage in the experiments it was decided to increase both the number of phonemes in the repertory of the machine and the number of words in the speech material to be used for testing. The phonemes /z/ and /f/ were added to the categories that the recogniser could deal with and a new, longer word list was made up. The extended list of words was obtained by selecting all those words found in a dictionary of about 60,000 common English words (35) which contained no phonemes other than the 13 in the repertory of the recogniser; as before, the long and short vowels /i:/ and /i/, and /ə:/ and /ə/ respectively were taken to be identical. This produced a list of just over 500 words which, of course, included many consonant clusters. The words were recorded and applied to the recogniser. Unfortunately it was found that the recognition of consonants in a cluster was very poor. This could have been remedied but only after extensive re-design of the acoustic recogniser and this was not considered worthwhile at this stage of the research. Just as an example of the kind of difficulty encountered, the phoneme /t/ was recognised by the characteristics of the fricative aspiration that follows the release of the stop in the articulation of the phoneme; when a /s/ follows a /t/ in a cluster, as for instance in the word *shoots*, the aspiration merges with the following fricative and cannot be detected. Another example of the difficulties encountered is when two plosives follow each other, as in the word *asked*. Here the /k/ is often not exploded and consequently only the aspiration of the following /t/ can be detected by the recogniser. In view of these difficulties it was decided to climinate from the list all words that contained consonant clusters. This resulted in list 3 which is given in Table 6 in the random order in which it was presented to subjects in some of the experiments to be described later. List 3 contains 200 words, of which 124 are monosyllabic, 70 disyllabic and 6 trisyllabic. The total number of sounds is 678 and the total number of digrams 878. The frequency of occurrence of the phonemes in the list is given in Table 7. The values of digram frequencies relevant to the words of List 3 are not given in this report because in all the experiments which used this word list, the linguistic store was adjusted to the so-called "pure CVCV" condition. This means that all CV and VC sequences were given equal probabilities whilst a CC or VV sequence was made impossible.

THE TESTING OF THE RECOGNISER

In the first series of experiments the words of List 1 were applied to the recogniser and the output was recorded by typewriter.

The effect of the linguistic information on the recognitions made by the machine was observed by reading the complete word list into the recogniser twice. On the first reading the stored knowledge of digram frequencies was not used and the output was determined solely by the acoustic recognition circuits; this mode of operation was available by making the probabilities of occurrence of all phonemes permanently equal and will be referred to in future as the unbiased condition. When the same word list was read into the machine for the second time, the store of linguistic knowledge was adjusted to give output voltages proportional to the digram frequencies, and the recogniser was said to operate in the biased condition. The potentiometer sliders of the linguistic store were adjusted to correspond to the digram frequency values given in Table 4 (a) by making the output corresponding to the largest digram frequency in any horizontal row in the table equal to 36 volts

Table 6

List of 200 words

List 3

car	mousse	teaser	litre
alarm	tars	meat	cur
marquees	tartars	looter	lucerne
curs	seek	catarrh	meek
loose	tartan	teasers	afoot
are	aloof	murmurs	neat
earl	fees	farm	cars
turf	myrrh	mars	two
Erse	lose	almoner	Caesar
coot	carcass	looser	meters
lunar	loom	shirkers	ark
feet	tar	cool	mazurka
cocoon	fern	feeler	salute
shoot	loser	occurs	anneal
machine	leak	loofah	tomb
zeal	mart	calf	firmer
litres	Turk	leas	mar
tarlatan	moot	sirs	shoes
sooner	earn	lark	feelers
leash	farmers	murmur	neater
carter	shirker	tool	saloon
tart	cart	soon	niece
art	mark	loot	knee
farmer	shark	mercer	combe
learners	accoutre	shoe	knees
fee	calm	meter	tarsus
seen	marsh	noon	learn

Table 6 (cont.)

assert	losers	mean	sees
eke	leaf	furs	lama
loon	lucre	psalm	surf
lamas	me	marquee	occur
turn	affirm	laugh	canoe
eel	Zulu	curt	teal
looters	furl	shirt	ooze
almoners	mercers	circus	team
afar	lean	zoo	ease
cartoon	cease	keys	sheet
shirk	meeker	term	sir
Zulus	fur	canoes	lea
lemurs	curse	sheen	martyr
see	teak	feel	eat
seat	lemur	tartar	farce
curl	alert	learner	meal
tarn	lurk	nurse	Shah
kneel	sheaf	martyrs	far
cooler	tease	arm	key
facetious	zoom	coon	moon
khan	lease	tea	salaam
firm	terse	armour	seem
noose	fool	keel	seal

Table 7

Frequency of occurrence of phonemes in word list 2 and word list 3.

Phoneme	Word List 2 Frequency of occurrence		Word List 3 Frequency of occurrence	
	Number of occurrences	Percentage of total no. of phonemes	Number of occurrences	Percentage of total no. of phonemes
t	36	13	61	9
k	29	11	51	8
s	26	10	40	6
ʃ	8	3	16	2
m	23	8	54	8
n	25	9	39	6
l	21	8	60	9
f	–	–	30	4
z	–	–	47	7
i:	31	11	63	9
a:	15	6	51	8
u:	16	6	49	7
ə:	40	15	117	17
Total	270	100	678	100

and all other outputs in the same row were made smaller proportionally to the appropriate digram frequency. For instance the phoneme /u:/ is followed most often, 30% of the time, by the phoneme /n/ and 11% of the time by /m/; therefore the potentiometer supplying information about the /u:/ to /n/ digram frequency is set to give an output of 36 volts and the one for the /u:/ to /m/ digram frequency an output of 13 volts. On the other hand, /i:/ is followed most frequently, 23% of the time, by /t/ and 15% of the time by /k/; therefore the voltage indicating the /i:/ to /t/ digram frequency is made 36 volts and that for /i:/ to /k/ is made 24 volts, and so on. Any digram frequency that would have had to be represented by an output of less than 5 volts – about 4 to 5% in most cases – was made to equal zero on the potentiometer matrix. The complete set of voltages from the linguistic store is given in Table 4(b).

Typical recognitions, as typed by the recogniser, are shown in Fig. 32.

seek	shooter	sheet
sik	5ute	5it
sik	5uta	5uns
sik	5uta	5it
meter	meeker	lease
mite	mike	lis
mitea	mtkea	is
mite	mike	l

Fig. 32. Typical recogniser outputs.
The first line gives the spelling of the word and the second line the phonemic transcription using the arbitrary symbols of the typewriter. The third and fourth lines show typical recogniser outputs for the unbiased and biased modes of operation respectively.

The actual characters printed by the typewriter could be chosen arbitrarily according to any desired code. The normal International Phonetic Association (I.P.A.) symbols were not used because the typewriter had a standard keyboard which did not include many of the accepted phonetic symbols. The phonemes /t, k, s, m, n, l/ were represented by the conventional symbols, the vowels /a:, i:, u: and ə:/were represented by the letters a, i, u, and e and the phoneme /ʃ/ was typed as the figure 5. In Fig. 32 the input word is shown in normal spelling in the top line and the transcription using the arbitrary symbols is shown in the second line; the second line then shows what the recogniser should type if it were working correctly. The actual output of the recogniser obtained without the use of the linguistic information is given in the third line and the output typed when it is using the stored information about digram frequencies is shown in the fourth and last line. The first word is an example of the case when the recogniser produces the desired output whether it uses linguistic information or not. The second word – *shooter* – is typed wrongly whatever the mode of operation of the recogniser: the linguistic information is apparently not strong enough to correct the final vowel wrongly recognised as /a:/ to an /ə:/ despite the fact that /t/ to /ə:/ digram frequency is more than three times as great as the one from /t/ to /a:/. The third word is a clear example of how the linguistic information can help in improving the performance of the recogniser: the /ʃ/ to /i:/ digram is about three times more probable than the /ʃ/ to /u:/ digram. The fourth and fifth words are additional examples of the same effect; they also show that the use of the linguistic information in many cases not only produces a correct recognition but also reduces the frequency with which additional symbols are typed, symbols that apparently have no counterpart in the speech input. The last example, the word *lease*, shows that the use of linguistic information can also have a detrimental effect: once the wrong recognition is made the digram frequencies produced for subsequent recognitions from the linguistic memory will also be wrong and the errors tend to be cumulative. In the case of the example, the acoustic input for the initial consonant /l/ was weak and similar in characteristics to the following vowel /i:/; when the recogniser was operating without linguistic information it

ignored the initial ambiguous segment and then typed the vowel and final consonant correctly. When the linguistic information was used it again missed the initial segment but recognised the vowel as an /1/ because initial vowels do not exist in its vocabulary. Once it had recognised the consonant, the linguistic information produced for the next recognition would exclude the possibility of another consonant; on the other hand the fricative characteristics of the next acoustic segment have no resemblance to any of the patterns stored for vowels. None of the final multiplication products will therefore be above the threshold set by the maximum detector and the only symbol typed for the whole word is the single *l*

THE PERFORMANCE OF THE RECOGNISER: INPUT/OUTPUT COMPARISONS

The performance of the recogniser was assessed by comparing the symbols typed at the output with the phoneme sequences of the input words; the phoneme recognitions, expressed as percentages of the total occurrence of each phoneme in the input, are shown in confusion matrices. Table 8 gives two sets of results, one obtained with the recogniser operating without the use of linguistic information and the other with the use of linguistic information. The column headed *Om.* gives the proportion of cases in which a particular phoneme at the input was omitted altogether from the output. Not shown in these matrices and not considered in the calculation of the percentages given in the table is the number of additional, unwanted symbols that were typed; the significance of these will be discussed later.

The overall score for assessing the performance of the recogniser was computed by considering three different kinds of error: incorrect recognitions (mistakes), omissions and extra symbols typed. The *error rate* was found to be 40% (or 60% correct recognitions) for the unbiased state and 28% (or 72% correct) for the biased state. The inclusion of the extra symbols typed in the computation of the error rates means that it is not really justifiable to deduce the score for correct recognitions from the error rates and they have therefore been given in brackets; when the extra symbols typed are not considered the score for correct recognitions becomes 72% and 75% for the unbiased and biased conditions respectively. Further information about the errors made by the recogniser is given in Table 9. The table gives separate figures for the total number of mistakes, omissions and extra symbols typed in both the unbiased and biased conditions and it also shows how these errors were distributed among the 11 phonemes in the recogniser's vocabulary. The number of mistakes, omissions and extra sounds typed are given as percentages of the total number of mistakes, etc. respectively; separate figures are given again for biased and unbiased operation. For example, in the unbiased state 10% of the total number of omissions (=21) were omissions of the phoneme /t/. Again, in the biased state 23% of the wrong recognitions (=43) were recognised as /m/, etc.

The results show that on the acoustic level the major difficulties were associated with the plosives and with the nasals and laterals. The plosives as a group could be identified quite readily, giving a score of about 85%, but many of the /k/ inputs were recognised as /t/, reducing the score for /k/ alone to 31%. Some experimental results obtained by the Haskins Laboratories in listening tests using synthetic speech may perhaps give some explanation of these difficulties. The automatic recogniser uses the spectrum of the "plosive burst" to distinguish between /t/ and /k/; the Haskins work has shown (39) that the spectrum of this burst for a particular plosive consonant varies as a function of the adjoining vowel; a high frequency "burst" clearly indicates a /t/ only when this burst is associated with an /i:/ vowel and it may be recognised more and more as a /k/ or /p/ as the vowel quality changes along the circumference of the vowel diagram through /a:/ to /u:/. Further, a medium frequency (around 1,500 c.p.s.) burst characterises a /k/ much more clearly when

Table 8

Confusion matrices for input/output comparisons, using word list 1.
Results are expressed as percentages of the total occurrences of each phoneme.

(a) Unbiased operation

Input \ Output	t	k	s	ʃ	m	n	l	i:	a:	u:	ə:	Om.
t	**84**		8							3		5
k	59	**31**										10
s			**96**									4
ʃ				**100**								
m					**83**	13						4
n					8	**76**	4					12
l					14		**24**	34		5	5	18
i:	3			3				**80**		3		11
a:					6				**87**			7
u:						6				**94**		
ə:				5				5	15		**67**	8

(b) Biased operation

Input \ Output	t	k	s	ʃ	m	n	l	i:	a:	u:	ə:	Om.
t	**86**		8			3						3
k	48	**31**					11					10
s			**92**									8
ʃ				**100**								
m					**87**	4	4					5
n					16	**60**	12					12
l					28		**67**					5
i:								**80**			7	13
a:									**100**			
u:										**88**		12
ə:								3	10		**64**	23

Table 9

Analysis of miatakes, omissions and extra aymbols typed by the machine in the biased (B) and in the unbiased (UB) mode of operation, uaing word list 2.

	Mistakes		Omissions		Extras	
	UB	B	UB	B	UB	B
Total number	54	43	21	26	35	5
	%	%	%	%	%	%
t	33	33	10	4	15	
k			14	12	9	20
s	6	7	5	8		20
ʃ	6				6	
m	11	23	5	4	3	
n	7	5	14	12	24	20
l	2	16	19	4	3	40
i:	17	2	14	15	9	
a:	11	9	5		15	
u:	6			8	9	
ə:	2	5	14	34	6	

aasociated with the back vowels like /a:/ and /u:/ than with a front vowel /i:/. The reaults obtained for /t/ and /k/ in a later series of experiments in which the words of Liat 3 were applied to the automatic recogniaer were apecially analyaed to aee how the correct recognition of these plosives was affected by the adjoining vowels. It was found that /t/ was always recognised correctly when adjoining the vowel /i:/, but only 70%, 50% and 25% of the time correctly when pronounced with the vowels /a:/, /ə:/ and /u:/ reapectively; similarly the /k/ recognitions were correct 70%, 60%, 50% and 20% of the time when the plosive was pronounced with /u:/, /a:/ /ə:/ and /i:/ respectively. Thia suggeats therefore that one way of getting better recognition of these plosives would have been to consider the nature of the adjoining vowel when asaessing the spectrum of the plosive burat.

As far as the nasals and laterals are concerned, even human listeners do not always find it easy to discriminate between /m/ and /n/ generally and between /m/, /n/ and /l/ when the discrimination has to be based on two formants only, as in the recogniser, instead of on three.

A comparison of the results obtained when the recogniser was operating in the biased and the unbiased condition indicates that although the use of linguistic information did not improve the score for correct phoneme recognitions to any great extent several significant differences can be observed between the two sets of scores. For instance the consonant /l/ was recognised correctly more than twice as often when the recogniser was working in the biased state than in the unbiased condition. Another difference between the two sets of results is the considerably smaller number of extra symbols typed in the biased condition, only 5 compared with 35 in the unbiased condition. Yet another difference is that the number of omissions is greater in the biased condition. This is largely due to cumulative errors occasionally produced by the use of linguistic information. A typical way in which this comes about, as has already been mentioned earlier, is that the machine fails to recognise the initial consonant of a word: when the acoustic recogniser afterwards, quite correctly, detects the following vowel the influence of the stored linguistic knowledge does not allow the vowel symbol to be typed and produces a consonant instead. Consequently during the next recognition the wrong set of digram frequencies will be utilised and either the wrong recognition is made again or none at all.

Despite the relatively small improvement in phoneme score and the increase in the number of omissions, the beneficial effect of using linguistic information is very noticeable. It is evident, even on first inspection, that the phoneme sequences typed when linguistic information is used are more like that of English and as a result, the words typed give the impression of being possible English words even if they do not make sense, whilst when no linguistic information is used many of the words typed have a distinctly non-English appearance like for instance the last but one example given in Fig. 32.

The word score, obtained by computing the proportion of complete words recognised without mistake, omission or extra phoneme, was then calculated for the results obtained in the biased and unbiased state of the recogniser. The results show that this word score has indeed increased considerably, from 24% in the unbiased state to 43% when linguistic information was used.

Although the comparison of input and output, in the way just described is useful for assessing the performance of the recogniser, it is by no means the only, and probably not even the most relevant way of deciding how far it is worthwhile using linguistic information in the automatic recognition process.

Before discussing the rationale of such other methods and the assessment of the recogniser's performance obtained by applying them, input/output comparisons for the results of some further experiments with the automatic recogniser will be described first. It will be remembered that a third word list, consisting of 200 words, was also prepared and that this list included two further phonemes, /f/ and /z/, in its vocabulary. An important reason for increasing the phoneme repertory was to extend the range of words that the recogniser could tackle: this was needed for making other experiments, to be described later, possible. The additional electronic circuits were put into use and the recorded word list was applied to the recogniser. Again, the recogniser was tested in the biased and unbiased condition. As with List 2, the difference between the overall phoneme scores obtained in the two modes

of operation is not very great: 62% in the unbiased condition and 68% when linguistic information is used. The number of extra characters typed, which were not included in the scores just quoted, was however considerably greater for both biased and unbiased operation. In the biased state the number of extra characters typed was 52, instead of 5 for List 2, and in the unbiased state 176. instead of 33 previously. This increase is partly due to the different, less careful, articulation of the speaker and partly due to the shortening of all time constants of the recogniser circuitry by about 20 to 25% which may well mean that some of the formant transitions are recognised as separate phonemes. As a result of the increased number of extra symbols typed the proportion of correctly recognised words has decreased to 35% in the biased state of the recogniser, as compared to 43% for List 2. A complete analysis of the output of the recogniser in its biased mode of operation is shown in Table 10.

As expected, the scores have not changed greatly as compared with List 2, except for the phonemes /t/, /f/, /s/ and /z/ which all use similar spectral cues. The additional errors in the recognition of these phonemes are largely within this group. A further change, as compared with List 2 is that all initial /m/, /n/ and /l/ sounds are grouped together under the label *m*, all final /m/ and /n/ sounds are labelled *n* and the final /l/ remains as *l*.

THE EFFECT OF USING MORE THAN ONE SPEAKER

It was obviously of interest to know how far the results achieved with one speaker's voice are maintained when the same words are spoken by a different speaker. The words of List 3 were spoken by two additional male speakers and the recordings used to test the recogniser. The voice of one speaker (F) was used in all experiments described so far and the circuitry was adjusted to perform best with his voice. The recordings made by the second and third speakers (T and G) were then applied one after the other to the recogniser operating in the biased mode. The results, shown in Table 11. indicate that the overall phoneme score has decreased to about 50% to 55% from the value of about 70% achieved with the voice of the first speaker and that the number of extra symbols typed has remained substantially unchanged. As a further test the circuitry of the recogniser was re-arranged to give the best possible results with the voice of the second speaker (T) rather than that of the first speaker (F). The re-arrangement consisted of connecting some of the multiplier inputs to different filters and of changes in the extent to which the filter output voltages were divided down before being applied to the multipliers. The scores obtained when the words spoken by (T) were now applied to the recogniser, both for the biased and unbiased modes of operation, are also shown in Table 11. It will be seen that the results obtained with this second voice are now very similar, both in terms of correctly recognised phonemes and of extra symbols typed, to those obtained with the first speaker's voice in the previous adjustment of the recogniser; the similarity is equally marked when the results in the biased mode are compared with each other and those in the unbiased mode. It seems then that the recogniser performs best when it is adjusted to the voice of one speaker and that the score drops markedly when another speaker is used; as far as one can tell from using only three different voices, this drop in score does not vary greatly from speaker to speaker. The fact that on using a different voice the recogniser's performance could be restored by relatively minor adjustments suggests that it would not be difficult to "teach" the recogniser to

Table 10

Confusion matrix for input/output comparison of results obtained in biased mode of recogniaer, using word list 3. Results are shown as percentages of the total occurrence of each phoneme.

Total phoneme score:	68	Conaonant score:	62
Vowel score:	77	Word score:	35

Input \ Output	a:	i:	u:	ə:	m	n	l	ʃ	t	k	f	s	z	Om.	Extra
a:	**79**		4	16										2	11
i:		**78**	17	3										1	13
u:	4	6	**75**	8										6	18
ə:	2	2	6	**76**										15	8
m					**86**	2	1	1		4	1			3	3
n					3	**62**	4	4		3	4			15	9
l						28	**61**	5						5	16
ʃ								**50**		37	12			0	12
t					2	2	3	3	**69**	7	11			3	2
k					2	2	10		35	**39**	8			4	6
f						3		3	20	13	**55**		6	0	6
s						2		2		2	2	**67**	17	5	2
z						13	6	2	6		8	2	**47**	15	0

Table 11

Comparison of results obtained when using the voices of three different speakers, F, T and G.

Speaker	Recogniser adjusted to deal optimally with voice shown below	Mode of operation of recogniser B/UB	Overall phoneme score %	Vowel score %	Consonant score %	No. of extra symbols typed
F	F	UB	62	54	68	176
F	F	B	68	77	62	52
T	F	B	51	66	41	51
G	F	B	56	61	52	62
T	T	UB	64	61	66	169
T	T	B	65	75	58	60

adjust itself to the voice of different speakers. One could arrange for example, that each fresh speaker would first have to say a test sentence; the recogniser, on being told that it is now dealing with a new voice and that the known test sentence is being spoken, would go through a pre-arranged routine of multiplier input changes, each time checking the degree of success. It could then choose that setting which gives the best performance in recognising the test sentence spoken with a fresh voice.

THE PERFORMANCE OF THE RECOGNISER: VISUAL AND ACOUSTIC TESTS

The comparison of input and output and the compiling of the confusion matrices based on these comparisons has proved a useful way of evaluating the recogniser, of finding the causes of errors and remedies for these. As has already been stated however, this may not be the only or necessarily the most relevant way of assessing the performance of a recogniser. Whatever the use to which the output of the recogniser is finally put, it will probably be presented in one form or another to a human "reader". This reader has to interpret, that is understand, the output and his own

knowledge of the language is available to correct some of the errors made by the automatic recogniser. The extent to which the reader can use his own linguistic knowledge will depend on the kind of mistakes made by the recogniser and also on the form in which the output is presented to him. The more familiar the form of presentation the easier the reader will find it to use his linguistic knowledge for this purpose and if the presentation is in an unfamiliar form then learning can make assimilation easier. Two new ways of assessing the performance of a recogniser and of the difference, if any, made by the use of linguistic information in the automatic recognition process then suggest themselves: one is to compare the reader's response to the output with the words applied to the input and the other is to find the amount of learning required by the reader in order to reach a given performance in understanding the recogniser's output.

As has just been mentioned, the way in which the reader can deal with the output of the recogniser depends on the form in which this output is presented to him and it seems worthwhile therefore to consider what are the most likely ways in which the output of the automatic recogniser will be put to practical use. Apart from its possible use for the voice control of machinery or of processes - an application where the question of a human reader does not arise anyway - the most likely applications are in analysis-synthesis telephony and as a speech typewriter. In the first of these applications, the phoneme sequence detected by the recogniser is transmitted and used to control some sort of speech synthesiser: the "reader" in this case will have to interpret an acoustic transform of a phoneme sequence or in other words he will deal with audible speech. In the second one of the above applications the output is presented to him in some form of writing which he has to read. The output of the recogniser was therefore presented in visual and acoustic form to separate groups of subjects to see how these forms of presentation compare and how far the reader can correct mistakes made by the recogniser. The words of List 2 were used and altogether 4 experiments were carried out: the output of the recogniser operating in the biased and in the unbiased mode was presented acoustically and visually to separate groups of subjects.

For acoustic presentation the phoneme sequence typed by the recogniser was pronounced by a speaker who was used to reproducing phonetic transcription. The reader produced the words on monotone and in the case of polysyllabic words equal stress was used for each syllable. The subjects were asked to write down, in normal spelling, whatever word they recognised.

In the case of visual presentation the output typed by the recogniser could have been used directly. The symbols typed for the different phonemes were, however, chosen at the time when the recogniser was constructed and with certain practical considerations in mind rather than from the point of view of what would be easy to read. The output was re-typed therefore, using a different set of symbols for the various phonemes, symbols which were as near as possible to normal English spelling and therefore it was thought could be read without difficulty by the average English speaking subject. The symbols used, together with key words, are shown in Table 12

Table 12

Key to the transliteration used in the visual presentation of the recogniser output.

The symbols used for the 11 phonemes are given side by side with key words to indicate their value.

t	t in **tool** or k in **cool**
k	k in **cool**
m	m in **mother** or n in **nothing** or l in **lesson**
n	n in **nothing**
l	l in **lesson**
s	s in **soak**
sh	sh in **shake** or in **sugar**
ee	ee in **fleet** or in **bean** (etc.)
oo	oo in **boot**
er	er in **burn** or in **after**
ah	ah in **barn** or in 2nd syllable of **shorter**

and were given in this form to all subjects in the visual experiments prior to the actual test. On inspecting this key it will be seen that some of the common mistakes made by the recogniser were also pointed out; for instance the recogniser frequently printed a *t* for a /k/ phoneme and therefore *tool* as well as *cool* are given as key words for *t*. The key gives alternative spellings for individual phonemes. For instance the key words given for /i:/ are *fleet* and *bean*; it was hoped to explain in this way that the symbols typed, *ee* in this case, represented phonemes rather than spelling forms. The subjects were given a sheet on which the output of the recogniser was printed in the transliteration of Table 12 and they were asked to write the words they recognised, in normal spelling alongside the appropriate printed transcription. Separate sheets were prepared for the outputs obtained when the recogniser was operating in the biased and unbiased mode and they were presented to different groups of subjects. As the subjects were asked to write down words, using normal spelling, a certain amount of care had to be exercised when marking the results so as to allow for the vagaries of spelling. For instance both *colonel* and *kernel* were taken as correct for the phoneme sequence /kə:nl/, both *loot* and *lute* for /lu:t/, etc. On the other hand *cease* was a correct response for the phoneme sequence /si:s/ but the word *sees* was an incorrect response.

Table 13

The performance of the recogniser in terms of input/output comparison and of the results of the visual and acoustic tests.

	No. of subjects UB	No. of subjects B	Unbiased: Sound score %	Unbiased: Word score %	Unbiased: Word score No.	Biased: Sound score %	Biased: Word score %	Biased: Word score No.	Words omitted UB %	Words omitted UB No.	Words omitted B %	Words omitted B No.
Input/output comparison			60	25	19	72	43	32				
Acoustic presentation:	33	33										
All words			52	30		64	46		18	454	15	371
Words recognised correctly by recogniser				84			84				3	33
Visual presentation:	21	20										
All words			41	28		57	43		40	630	27	405
Words recognised correctly by recogniser				88			81				7	42

Table 14

Confusion matrices obtained from the responses to visual (a) and acoustic (b) presentations of the biased recogniser output. Confusions are expressed as a percentage of the total occurrences of each phoneme; those amounting to less than 1 per cent are disregarded. The column headed Om. gives the percentage of cases in which the phoneme was either omitted or replaced by some phoneme outside the repertory of the machine.

(a) **Visual**

Phoneme presented \ Response	t	k	s	ʃ	m	n	l	i:	a:	u:	ə:	Om.
t	**65**	8	2									25
k	16	**44**				1	2					37
s			**56**									44
ʃ				**80**								20
m					**56**	13	7					24
n					4	**51**	6					39
l					15	3	**41**					40
i:							2	**68**			3	27
a:									**65**			35
u:										**52**		48
ə:											**46**	53

(b) **Acoustic**

Phoneme presented \ Response	t	k	s	ʃ	m	n	l	i:	a:	u:	ə:	Om.
t	**67**	8	4									20
k	21	**48**			2	2	3					24
s			**67**				1					31
ʃ			2	**82**								15
m					**63**	15	6					16
n					5	**59**	10					26
l					15	3	**51**				2	29
i:							2	**75**			3	20
a:									**78**			22
u:								2		**63**		35
ə:								2	2	1	**51**	44

The results of the visual and of the acoustic tests are summarised in Table 13 and confusion matrices for the results obtsined from the visual and acoustic presentation of the biased output of the recogniser are shown in Table 14. The data show that the results for presenting the output of the recogniser to human interpreters are similar to those for input/output comparison; also the overall scores from the responses to acoustic presentation are not very different from those from the visual presentation. When subjects' responses are examined in more detail, however, a number of significant differences between responses to the two modes of presentation can be observed, showing that in effect the subjects go through a somewhat different recognition process in the two cases. In the normal process of speech recognition the subject is used to interpreting sound patterns with reference to his linguistic memory and therefore when the recogniser's output is presented to him acoustically he can use this memory directly. On the other hand when the recogniser's output is given to him visually he first has to go through a process of thinking of the acoustic form of the printed symbols before he can use his linguistic memory. It seems that the subjects did not find this an easy process, although they all had previous experience of reading phonetic transcription. For example the words *colonel*, *sooner* and *circle*, because of mistakes made by the recogniser, sppesred in the visual presentation as *lernl*, *soomerl* and *serker*. None of the subjects in the visual tests recognised theae words correctly but about 25% of those in the acoustic tests did so.

It seems also thst subjects doing the scoustic tests were much more likely to make phonemic substitutions whilst those presented with the visual form of the output tended to operate with complete words. One indicstion for this is that, when in doubt, the subjects in the acoustic tests experimented freely with phonemic substitutions in order to produce a word that they thought might be the right answer whilst those doing the visual tests often did not make a response at sll under these circumstances. It was perhaps as s result of this thst the number of omissions of complete words in the two acoustic presentations amounted to only 18% and 15% of the total number of words whilst the corresponding figures for the visual tests were 40% and 27%. This tendency could be observed even for the words thst were typed correctly by the sutomatic recogniser: only 3% were omitted entirely in the acoustic presentation ss compared with about 7% in the visusl test.

Evidence for the greater facility of making phonemic substitutions in the scoustic form of presentation is the response made by subjects to words correctly typed by the recogniser. For example, the word *tool* was typed by the recogniser without mistake so that in the acoustic test it was hesrd correctly and the convention for trsnsliteration wss such thst the word was even presented with the correct spelling in the visual test. Nevertheless, only sbout 60% of the subjects in the acoustic test recognised the word correctly, most of the others substituting the word *cool*, whilst in the visusl test 90% responded correctly.

The results sre, of course, affected by s number of other fsctors. For example the words used are not very homogeneous: words that phoneticslly or in spelling are quite close to each other might differ greatly in their frequency of occurrence in the language and therefore in the extent to which subjects expect them. For instance, in the scoustic test the word *meeker*, correctly typed by the recogniser, was identified correctly only 43% of the time and 28% of the time as meter, whilst the word *meter*, also correctly typed by the recogniser was identified 82% of the time; in the visual tests both *meeker* and *meter* were recognised correctly only sbout 55% of the time.

THE INFLUENCE OF CONTEXT ON SUBJECTS' ABILITY TO INTERPRET THE OUTPUT OF THE RECOGNISER

The last experiment to be described concerned the effect of the subjects' expectations on their ability to interpret the output of the recogniser: altering these expectations is one way of changing the linguistic constraints which affected the subject. As there was a definite limit to the amount of constraint that could be included in the machine it was of great interest to know how far variations of the constraints affecting the "reader" altered the overall performance.

The words of List 3 were used in the experiment; the presentation was always in the visual form and the normal I.P.A. characters were used, instead of the ones in the previous tests, because all 18 subjects were quite fluent in the use of these symbols. Four separate lists of words were used. The first one was the entire List 3; the other three lists consisted of words selected from List 3, words whose meaning had something in common. The words of one list all had something to do with water, of the other one with humour and pastimes and those of the third list were all adjectives. The actual words in these three lists are given in Table 15.

The output of the recogniser for the words of List 3 was obtained first. This output was then presented to the subjects in separate tests, first for all the words of List 3 and then for the words of the selections of Table 15 in turn. They were told that the first list was a general one, whilst the meaning of the words of the other lists had something in common as shown by the table headings and they were asked to write down in ordinary spelling what they thought the words were. The responses to the last three lists were then scored for correct recognition and this score was compared with that obtained *for the same words* in the general list. The results show that the scores always improved when the extra contextual clue was available but the improvement was only marginal for the words connected with water, 25% to 27%, and the words connected with humour and pastimes, 25% to 26%. The improvement was more noticeable, 19% to 30%, for the list containing adjectives. The results must naturally be highly dependent on the closeness of the common meaning of the words in any one list and the scope of the words in the vocabulary of the recogniser (List 3) was not extensive enough to make up really satisfactory groups of words.

The question of the effect of the human termination on the operation of an automatic speech recogniser is obviously an important one and much further work is needed to investigate it.

Table 15

Lists of words whose meanings have something in common. The words of all three lists are included in word list 3.

(a) Words to do with water	(b) Words to do with humour and pastimes	(c) Adjectives, including participles, excluding nouns that may be used attributively
calm	mazurka	neat
keel	art	lean
ark	circus	Zulu
canoes	teasers	firmer
sea	farce	Erse
eel	facetious	neater
leak	shoot	calm
coot	turn	firm
teal	cartoon	loose
ooze	fool	seen
marsh	turf	far
canoe	laugh	two
teem	teaser	terse
surf	saloon	cooler
shark	tease	cool
seas		looser
tarn		meek
seal		lunar
		mean
		aloof
		meeker
		curt
		tart
		facetious

CHAPTER VI

CONCLUSIONS

On reviewing the work described in this report it can be stated that an automatic speech recogniser, utilising both acoustic and linguistic information in its recognition processes, has been constructed. The circuitry dealing with the recognition of the acoustic characteristics searches for the presence of well-known acoustic correlates of the phonemes and provides information about the probability of occurrence of these phonemes from the acoustic point of view. The store of linguistic knowledge provides an estimate of the linguistic probability of the occurrence of the phonemes. The recogniser selects that phoneme for which the combined probabilities are greatest. The recogniser can deal with altogether 13 phonemes; 9 consonants and 4 vowels. The principal aim of the experiments was to investigate how far the use of linguistic information improves the performance of the recognition process. Some additional experiments were carried out to see how far the linguistic knowledge of a human observer can be used to improve the performance of a recogniser when he is asked to interpret its output.

The results of the experiments show that the use of even a very limited amount of linguistic information does help in the recognition process: some phoneme sequences impossible in English were eliminated from the output and the overall word score improved by 50% from 28% to 43%. The results also show however, that the use of linguistic information can make the results worse as well as better: once a mistake has been made, the wrong kind of linguistic information is utilised and a further error is made that might have been avoided had linguistic information not been used. The detrimental effect of this procedure was minimised by restricting the speech material to words spoken in isolation and the silence between words was used to check, at frequent intervals, that the correct set of digram frequencies was being utilised. A more fundamental way of rectifying this kind of error and also of improving the performance of the recogniser is to organise the linguistic store on several levels, the phonemic and the word levels for example. This can be understood best by comparing two automatic recognition systems, one in which only phonemes and phoneme sequential probabilities to n places are stored, and the other which also remembers the words of the language it deals with, the words being stored as sequences of phonemes up to n places. In the system which operates solely with phonemes, successive recognitions are made in the light of preceding phonemes only and once made cannot be corrected; if an error is made it will prejudice all future recognitions. In the other system a whole sequence of phonemes is recognised only provisionally at first and the sequence is compared with the word store to find a best match. The final decision is then reached in the light of the following as well as of the preceding phonemes. A further advantage of the second system is that the output is necessarily in the form of words whilst the purely phonemic system can produce phoneme sequences that do not form meaningful words. Once a multi-level system of the kind just described has been established, its performance can be improved further by making the phoneme sequential probabilities dependent upon preceding word recognitions. Such feed-back of information from level to level can increase the constraints considerably.

Phonemes and words are, of course, not the only levels of linguistic organisation that can be included in the linguistic knowledge of a recogniser. Further linguistic knowledge based on word transition probabilities and on a sentence store would also improve the performance. The use of such knowledge, particularly of an adequate sentence store, would require very considerable storage capacity. A more modest, though necessarily less all-embracing, way of using sentence information is to store enough information about sentence structure for enabling the machine to recognise the syntactical elements of the input sentence and then to modify constraints on word and phoneme levels accordingly.

The larger linguistic units need not be stored solely as sequences of the smaller units. Acoustic patterns corresponding to the larger units could also form the basis of a recognition process. This would mean recognition in terms of the longer acoustic sequences that are stored for the larger linguistic units and it can be expected that this would offer an advantage over the recognition of a long sequence of shorter units as used for phonemic recognition, because of the acoustic or articulatory constraints operative in speech.

This latter question is just a small detail of the much larger problem of deciding whether it is more rewarding to improve the sophistication of the acoustic recogniser or to extend the linguistic knowledge of the machine. This question can only be decided on empirical rather than on theoretical grounds. As work on automatic speech recognition progresses, practical systems using these alternative principles of recognition and giving comparable levels of performance must be compared to see which one offers greater economy of instrumentation.

So far in the discussion it has always been assumed that the only way of increasing the linguistic constraints effective in the recognition process is to augment the linguistic knowledge stored in the machine. In fact, the linguistic constraints can also be increased by restricting the variety of the speech material used as the input to the recogniser. Depending on the way such restrictions are applied, they can increase the constraints either in the machine or in the human "reader" of the output. A few experiments on the use of the latter of these possibilities have already been described in this report. As far as the former method is concerned, a more restricted speech input will simplify the task of the machine not only because of the smaller number of choices offered but also because by suitable selection the variety of phonemes and of phoneme transition probabilities can also be reduced. An example of this is the restriction of the possible words to those which do not contain consonant clusters. Some indication of how such factors operate has been given in the experiments described above: when the repertory of phonemes was increased from 11 to 13 the phoneme score fell from 75% to 68% and the word score from 43% to 35%.

It seems likely that automatic recognisers designed for a considerably restricted speech material will acquire practical importance: experimental results suggest that whilst not enough is known yet to make recognisers dealing with English speech generally a practical possibility, it is possible that a recogniser designed to identify a small number, say up to 30 or 50, words spoken in isolation may operate successfully in the none too distant future. It may well be that many of the methods used in such a specialised recogniser offer a solution that is relevant only to the restricted input condition; it is equally likely though that such machines will also produce some pointers useful for the solution of the general problem. Therefore the design of such restricted machines should be interesting from the theoretical as well as from the practical point of view.

Most future experiments on automatic speech recognition will probably require the storage of considerable amounts of information, the selective use of such information and the making of decisions dependent on a variety of contingencies. The large digital computers available commercially offer such facilities. These computers are also suitable for collecting much of the statistical information, about both the acoustic and linguistic aspects of speech, that are needed for various automatic speech recognition processes and some of which are detailed in another publication (25). It is likely therefore that computers will find considerable application in this branch of speech research. Although they are expensive to rent, the alternative method of constructing specialised circuitry to perform these operations would be even more expensive and time consuming. It is hoped that by using computers a variety of automatic recognition processes can be tried out, evaluated and compared in a relatively short time. Whenever a method of recognition of practical importance has been found it should not be too difficult to transform the computer programme into a practical electronic circuit performing the same function. Work on finding the best ways of using computers for research on speech and automatic speech recognition is already in progress.

REFERENCES

1. Ayers, E.L., An analysis-synthesis telephone system. *Report of Colloquium at Signals Research & Development Establishment.* S.R.D.E. Report No. 1100, 1956, 28-32.

2. Brown, T.L. and Stephenson, W.L., A stabilised DC power supply using transistors. *Electronic Eng.*, 29 (1957), 425-8.

3. Cooper, F.S., Liberman, A.M. and Borst J.M., The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proc. Nat. Acad. Sc.*, 37 (1951), 318-25.

4. David, E.E., Signal theory in speech transmission. *Inst. Radio Enginrs. Trans. on Circuit Theory*, Vol. CT-3 (Dec. 1956), 232-44.

5. Denes, P., Effect of duration on the perception of voicing. *J. acoust. Soc. Amer.*, 27 (1955), 761-4.

6. Dolansky, L.O., An instantaneous pitch period indicator. *J. acoust. Soc. Amer.*, 27 (1955), 67-72.

7. Dreyfus-Graf, J., Sonograph and sound mechanics. *J. acoust. Soc. Amer.*, 22 (1950), 731-9.

8. Dreyfus-Graf, J., Phonetographe et subformants. *Bull. Technique PTT*, No. 2, (1957), 4-59.

9. Dudley, H., Remaking speech. *J. acoust. Soc. Amer.*, 11 (1939), 169-77.

10. Dudley, H., The carrier nature of speech. *Bell Syst. Tech. J.*, 19 (1940), 495-515.

11. Dudley, H. and Balashek, S., Automatic recognition of phonetic patterns in speech. *J. acoust. Soc. Amer.*, 30 (1958), 721-32.

12. Dudley, H., Phonetic pattern recognition vocoder for narrow band speech transmission. *J. acoust. Soc. Amer.*, 30 (1958), 733-9.

13. Dunn, H.K., On vowel resonances and an electrical vocal tract. *J. acoust. Soc. Amer.*, 22 (1950), 740-53.

14. Fant, C.G.M., On the predictability of formant levels and spectrum envelopes from formant frequencies. *For Roman Jakobson*, 109-20. The Hague: Mouton, 1956.

15. Fant, C.G.M., *Acoustic Theory of Speech Production.* Royal Institute of Technology, Stockholm, Report No. 10, 1958.

16. Flanagan, J.L., Difference limen for vowel formant frequency. *J. acoust. Soc. Amer.*, 27 (1955), 613-7.

17. Flanagan, J.L., Difference limen for the intensity of a vowel sound. *J. acoust. Soc. Amer.*, 27 (1955), 1223-5.

18. Flanagan, J.L. and House, A.S., Development and testing of a formant-coding speech transmission system. *J. acoust. Soc. Amer.*, 28 (1956), 1099-1106.

19. Flanagan, J.L. and Saslow, M.G., Pitch discrimination for synthetic vowels. *J. acoust. Soc. Amer.*, 30 (1958), 435-42.

20. Flanagan, J.L., Resonance vocoder and baseband complement. *Inst. Radio Eng. Wescon Convention Record*, Vol. 3, Pt. 7 (Aug. 1959), 5-16.

21. Fletcher, H., *Speech and Hearing*. New York: Van Nostrand, 1929.

22. Fry, D.B., The experimental study of speech. *Studies In Communication*, 147-67. London: Secker & Warburg, 1955.

23. Fry, D.B. and Denes, P., On presenting the output of a mechanical speech recogniser. *J. acoust. Soc. Amer.*, 29 (1957), 364-7.

24. Fry, D.B. and Denes, P., The solution of some fundamental problems in mechanical speech recognition. *Language and Speech*, 1 (1958), 35-58.

25. Fry, D.B. and Denes, P., An analogue of the speech recognition process. *Mechanization of Thought Processes: National Physical Laboratory Symposium No. 10*, 375-84. London: H.M. Stationery Office, 1959.

26. Gabor, D., Theory of communication. *J. Inst. Elec. Engnrs*, 93 (1946), Pt.III, 429-57.

27. Gill, J.S., Automatic extraction of the excitation function of speech with particular reference to the use of correlation methods. Paper given at 3rd Intern. Congress on Acoustics, Stuttgart, 1959. To be published.

28. Gruenz, O.O. and Schott, L.O., Extraction and portrayal of pitch of speech sounds. *J. acoust. Soc. Amer.*, 21 (1949), 487-95.

29. Halsey, R.J. and Swaffield, J., Analysis-synthesis telephony, with special reference to the vocoder. *J. Inst. Elec. Engnrs*, 95 (1948), Pt. III, 391-406.

30. Harris, K.S., Cues for the discrimination of American English fricatives in spoken syllables. *Language and Speech*, 1 (1958), 1-7.

31. Heinz, J.M., Model studies of the production of fricative consonants. *Quarterly Progress Reports*, 1956-58, Research Laboratory of Electronics, Massachusetts Institute of Technology.

32. Howard, C.R., Speech analysis-synthesis scheme using continuous parameters. *J. acoust. Soc. Amer.*, 28 (1956), 1091-8.

33. Hughes, G.W. and Halle, M., Spectral properties of fricative consonants. *J. acoust. Soc. Amer.*, 28 (1956), 303-10.

34. Jakobson, R., Fant, C.G.M. and Halle, M., *Preliminaries to speech analysis*. Technical Report No. 13, Acoustics Laboratory, Massachusetts Institute of Technology, 1952.

35. Jones, D., *An English Pronouncing Dictionary*. London: J.M. Dent, 1956.

36. Koenig, W., A new frequency scale for acoustic measurements. *Bell Lab. Record*, 27 (1949), 299-301.

37. Ladefoged, P. and Broadbent, D.E., Information conveyed by vowels. *J. acoust. Soc. Amer.*, 29 (1957), 98-104.

38. Lawrence, W., Synthesis of speech from signals which have a low information rate. *Communication Theory*, 460. London: Butterworth, 1953.

39. Liberman, A.M., Delattre, P., and Cooper, F.S., The role of selected stimulus variables in the perception of the unvoiced stop consonants. *Amer. J. Psychol.*, 65 (1952), 497-516.

40. Liberman, A.M., Delattre, P.C., Cooper, F.S., and Gerstman, L.J., The role of consonant-vowel transitions in the perception of stop and nasal consonants. *Psychol. Mono.*, 68 (1954), No. 1, 1-13.

41. Liberman, A.M., Some results of research on speech perception. *J. acoust. Soc. Amer.*, 29 (1957), 117-23.

42. Liberman, A.M., Ingemann, F., Lisker, L., Delattre, P. and Cooper, F.S. Minimal rules for synthesising speech. *J. acoust. Soc. Amer.*, 31 (1959), 1490-9.

43. O'Connor, J.D., Gerstman, L.J., Liberman, A.M., Delattre, P. and Cooper, F.S., Acoustic cues for the perception of initial /w j r l/ in English. *Word*, 13 (1957), 24-43.

44. Peterson, G.E., and Barney, H.L., Control methods used in a study of the vowels. *J. acoust. Soc. Amer.*, 24 (1952), 175-84.

45. Potter, R.K., Kopp, G.A. and Green, H.C., *Visible Speech*. New York: Van Nostrand, 1947.

46. Reynolds, R.E., A regulated power unit with transistor control. *Mullard Tech. Communications*, 3 (1957), 34-6.

47. Schroeder, M.R. and David, E.E., A vocoder for transmitting 10 Kc.p.s. speech over a 3.5 Kc.p.s. channel. *Acustica*. To be published.

48. Shannon, C.E. and Weaver, W., *The Mathematical Theory of Communication*. Urbana: University of Illinois Press, 1949.

49. Smith, C.P., Speech data reduction. Ref. AFCRC-TR-57-111. ASTIA Document No. AD 117290, 1957.

50. Stead, L.G. and Jones, E.T., The S.R.D.E. speech bandwidth compression project. *Proc. Seminar on Speech Compression and Processing*, Sept. 1959. Air Force Cambridge Research Center, Mass. Report No. AFCRC-TR-59-198, Vol. 1.

51. Stevens, K.N. and House, A.S., Development of a quantitative description of vowel articulation. *J. acoust. Soc. Amer.*, 27 (1955), 484-93.

52. Stevens, K.N. and House, A.S., Studies of formant transitions using a vocal tract analog. *J. acoust. Soc. Amer.*, 28 (1956), 578-85.

53. Strevens, P., Spectra of fricative noise in human speech. *Language and Speech*, 3 (1960), 32-49.

54. Swaffield, J., Some progress with vocoder type systems. *Proc. Seminar on Speech Compression and Processing*, Sept. 1959. Air Force Cambridge Research Center, Mass. Report No. AFCRC-TR-59-198, Vol. 1.

55. Wiren, J. and Stubbs, H.L., Electronic binary selection system for phoneme classification. *J. acoust. Soc. Amer.*, 28 (1956), 1082-91.